test for class

lend jwj cen

2018-11-19

Contents

Chapter 1

簡介

測試這個無言的語言

無關紀錄先前的設定 tango yaml 不能用中文? 測試紀錄, 簡單提要, 各種其他文章複製參考

1.1 網路資源

- 可以參考
- rmd cheat sheet
- gitbook
- pimp my rmd
- ENDMEMO
- R Package Primer

1.2 安裝軟體

部份軟體列表另外於 linux 安裝 tidyverse 時需要:

```
sudo apt-get install -y libxml2-dev libcurl4-openssl-dev libssl-dev
```

install.packages()

```
install.packages('devtools')
install.packages('bookdown')
install.packages('rlang')
install.packages('tidyr')
install.packages('babynames')
install.packages('ggplot2')
install.packages('sm')
system('sudo apt-get install -y libxml2-dev libcurl4-openssl-dev libssl-dev')
install.packages('tidyverse')
install.packages('codetools')
install.packages('moments')
##
options(repos = c(CRAN = "https://cran.revolutionanalytics.com"))
```

1.2.1 自動 update package

```
all.packages <- installed.packages()
r.version <- paste(version[['major']], '.', version[['minor']], sep = '')

for (i in 1:nrow(all.packages))
{
    package.name <- all.packages[i, 1]
    package.version <- all.packages[i, 3]
    if (package.version != r.version)
    {
        print(paste('Installing', package.name))
        install.packages(package.name)
    }
}
```

1.3 quick view

快速指令

1.3.0.1 目前有哪些資料集可以測試

```
data()
```

1.4 資料型態和內容

可以先看看資料描述 ?mtcars

```
mtcars
```

```
                     mpg cyl  disp  hp drat   wt qsec vs am gear
Mazda RX4           21.0   6 160.0 110 3.90 2.62 16.5  0  1    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.88 17.0  0  1    4
Datsun 710          22.8   4 108.0  93 3.85 2.32 18.6  1  1    4
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.21 19.4  1  0    3
Hornet Sportabout   18.7   8 360.0 175 3.15 3.44 17.0  0  0    3
Valiant             18.1   6 225.0 105 2.76 3.46 20.2  1  0    3
Duster 360          14.3   8 360.0 245 3.21 3.57 15.8  0  0    3
Merc 240D           24.4   4 146.7  62 3.69 3.19 20.0  1  0    4
Merc 230            22.8   4 140.8  95 3.92 3.15 22.9  1  0    4
Merc 280            19.2   6 167.6 123 3.92 3.44 18.3  1  0    4
Merc 280C           17.8   6 167.6 123 3.92 3.44 18.9  1  0    4
Merc 450SE          16.4   8 275.8 180 3.07 4.07 17.4  0  0    3
Merc 450SL          17.3   8 275.8 180 3.07 3.73 17.6  0  0    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.78 18.0  0  0    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.25 18.0  0  0    3
Lincoln Continental 10.4   8 460.0 215 3.00 5.42 17.8  0  0    3
Chrysler Imperial   14.7   8 440.0 230 3.23 5.34 17.4  0  0    3
Fiat 128            32.4   4  78.7  66 4.08 2.20 19.5  1  1    4
Honda Civic         30.4   4  75.7  52 4.93 1.61 18.5  1  1    4
Toyota Corolla      33.9   4  71.1  65 4.22 1.83 19.9  1  1    4
Toyota Corona       21.5   4 120.1  97 3.70 2.46 20.0  1  0    3
Dodge Challenger    15.5   8 318.0 150 2.76 3.52 16.9  0  0    3
AMC Javelin         15.2   8 304.0 150 3.15 3.44 17.3  0  0    3
Camaro Z28          13.3   8 350.0 245 3.73 3.84 15.4  0  0    3
Pontiac Firebird    19.2   8 400.0 175 3.08 3.85 17.1  0  0    3
Fiat X1-9           27.3   4  79.0  66 4.08 1.94 18.9  1  1    4
Porsche 914-2       26.0   4 120.3  91 4.43 2.14 16.7  0  1    5
Lotus Europa        30.4   4  95.1 113 3.77 1.51 16.9  1  1    5
Ford Pantera L      15.8   8 351.0 264 4.22 3.17 14.5  0  1    5
Ferrari Dino        19.7   6 145.0 175 3.62 2.77 15.5  0  1    5
Maserati Bora       15.0   8 301.0 335 3.54 3.57 14.6  0  1    5
Volvo 142E          21.4   4 121.0 109 4.11 2.78 18.6  1  1    4
                    carb
Mazda RX4              4
Mazda RX4 Wag          4
Datsun 710             1
Hornet 4 Drive         1
Hornet Sportabout      2
Valiant                1
Duster 360             4
Merc 240D              2
Merc 230               2
Merc 280               4
Merc 280C              4
Merc 450SE             3
Merc 450SL             3
Merc 450SLC            3
Cadillac Fleetwood     4
Lincoln Continental    4
Chrysler Imperial      4
Fiat 128               1
Honda Civic            2
Toyota Corolla         1
Toyota Corona          1
Dodge Challenger       2
AMC Javelin            2
Camaro Z28             4
Pontiac Firebird       2
Fiat X1-9              1
Porsche 914-2          2
Lotus Europa           2
Ford Pantera L         4
Ferrari Dino           6
Maserati Bora          8
Volvo 142E             2
```

```
head(mtcars)
```

```
                   mpg cyl disp  hp drat   wt qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.62 16.5  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.88 17.0  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.32 18.6  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.21 19.4  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.44 17.0  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.46 20.2  1  0    3    1
```

```
tail(mtcars)
```

```
                mpg cyl  disp  hp drat   wt qsec vs am gear carb
Porsche 914-2  26.0   4 120.3  91 4.43 2.14 16.7  0  1    5    2
Lotus Europa   30.4   4  95.1 113 3.77 1.51 16.9  1  1    5    2
Ford Pantera L 15.8   8 351.0 264 4.22 3.17 14.5  0  1    5    4
Ferrari Dino   19.7   6 145.0 175 3.62 2.77 15.5  0  1    5    6
Maserati Bora  15.0   8 301.0 335 3.54 3.57 14.6  0  1    5    8
Volvo 142E     21.4   4 121.0 109 4.11 2.78 18.6  1  1    4    2
```

1.4.1 編輯/瀏覽資料

```
edit(mtcars)
data.entry(mtcars)
View(mtcars)
```

1.4.2 個別欄位

如果要顯示個別欄位, 一般可以是 mtcars$mpg, 但是如果要直接使用 mpg 欄位, 可以利用 attach()

```
attach(mtcars)
mpg
```

```
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3
[14] 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3
[27] 26.0 30.4 15.8 19.7 15.0 21.4
```

1.4.3 質性數據的分析

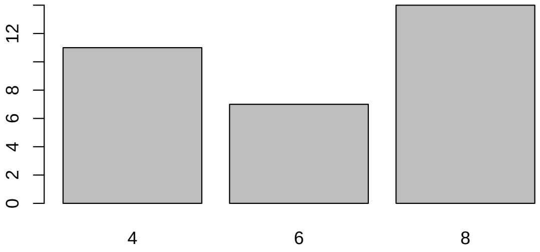
欄位 cyl 為質性變數, 可以利用 table 分析

```
table(mtcars$cyl)
```

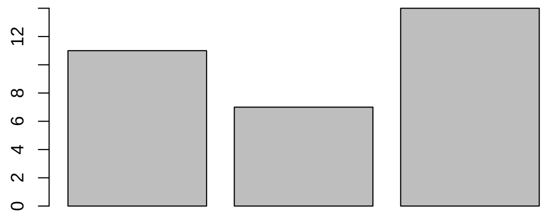
```

 4  6  8 
11  7 14 
```

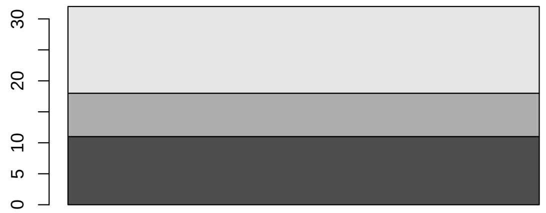
頻率圖

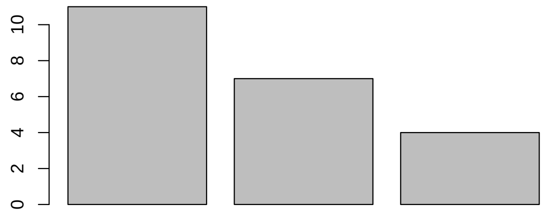
```
barplot(table(mtcars$cyl))
```



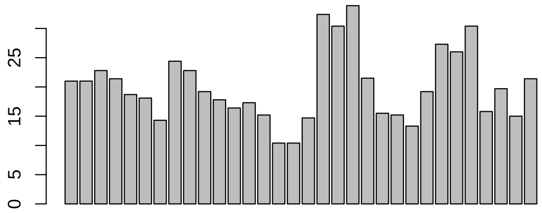
```
v<-as.vector(table(mtcars$cyl))
barplot(v)
m<-as.matrix(table(mtcars$cyl))
barplot(m)

barplot(c(11,7,4))
```

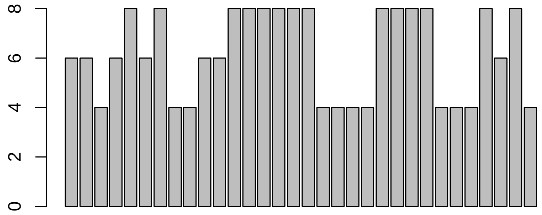






```
barplot(mtcars$mpg)
barplot(mtcars$cyl)
```





1.5 assignment

<-和 -> 是一對，可以向左和向右賦值

= 是單向的，作用和 <-基本相同，但對函數中的變數通常使用 =

<<- 這個是全域賦值，跟變數的作用域有關，一般不會用到

```
##Delete x (if it exists)
rm(x)
```

```
Warning in rm(x): 找不到物件 'x'
```

```
mean(x = 1:10) #[1] 5.5
```

```
[1] 5.5
```

```
x #Error: object 'x' not found
```

```
Error in eval(expr, envir, enclos): 
  找不到物件 'x'
```

Here x is declared within the function 's scope of the function, so it doesn 't exist in the user workspace. Now, let 's run the same piece of code with using the <- operator:

```
mean(x <- 1:10) # [1] 5.5
```

```
[1] 5.5
```

```
x
```

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

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

This time the x variable is declared within the user workspace. When does the assignment take place ? In the code above, you may be tempted to thing that we "assign 1:10 to x, then calculate the mean." This would be true for #languages such as C, but it isn 't true in R. Consider the following function:

```
a <- 1
f <- function(a) return(TRUE)
f <- f(a <- a + 1);
# 輸出:TRUE
a # 結果 =1
```

```
[1] 1
```

Notice that the value of a hasn 't changed! In R, the value of a will only change if we need to evaluate the argument in the function. This can lead to unpredictable behaviour:

```
f <- function(a) if (runif(1) > 0.5) TRUE else a
    f(a <- a + 1);
a # result 2
```

```
[1] 2
```

```
f(a <- a + 1);
```

```
[1] TRUE
```

```
# TRUE
a # 2
```

```
[1] 2
```

```
f(a <- a + 1);
```

```
[1] TRUE
```

```
a #3
```

```
[1] 2
```

= 用在參數指派例如

```
matrix(1:20, ncol = 4)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

如果

```
matrix(1:20, ncol <- 4)
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    5    9   13   17
[2,]    2    6   10   14   18
[3,]    3    7   11   15   19
[4,]    4    8   12   16   20
```

```
ncol
```

```
[1] 4
```

會產生一個變數 ncol 結論:x<-3 <-會在全局產生變數 x 然後指派 3

```
(x <- 3)
```

```
[1] 3
```

```
#rm(list = ls())
rm(x)
ls()
```

```
[1] "a"    "f"    "m"    "ncol" "v"
```

```
(x = 3)
```

```
[1] 3
```

```
ls()
```

```
[1] "a"    "f"    "m"    "ncol" "v"    "x"
```

因為 x 是參數名稱不是變數, 看 mean help

Chapter 2

data type

2.1 基本操作

2.1.1 指派

雖然也可以用 = 但是，R 的設計是使用 <-。

```
a <- 3
```

a<-3 是一個指派的敘述句, 不會回饋資訊到螢幕上。如果要知道 a 的內容是甚麼, 就打入 a 或者 (a<-3)

```
    a
```

```
[1] 3
```

```
(a<-3)
```

```
[1] 3
```

```
b <- sqrt(a * a + 5)
b
```

```
[1] 3.74
```

在 session 中的如果要列出已經定義過的變數, 可以利用 ls

```
ls()
```

```
[1] "a" "b"
```

2.2 運算

2.2.1 四則運算

2.2.2 Logical

Another important data type is the logical type. There are two predefined variables, TRUE and FALSE:

```
a = TRUE
typeof(a)
```

```
[1] "logical"
```

```
b = FALSE
typeof(b)
```

```
[1] "logical"
```

The standard logical operators can be used:

operator	說明
<	less than
>	great than
<=	less than or equal
>=	greater than or equal
==	equal to
!=	not equal to
|	entry wise or
||	or
!	not
&	entry wise and
&&	and
xor(a,b)	exclusive or

Note that there is a difference between operators that act on entries within a vector and the whole vector:

```
a = c(TRUE, FALSE)
b = c(FALSE, FALSE)
a | b
```

```
[1]  TRUE FALSE
```

```
a || b
```

```
[1] TRUE
```

```
xor(a, b)
```

```
[1]  TRUE FALSE
```

There are a large number of functions that test to determine the type of a variable. For example the is.numeric function can determine if a variable is numeric:

```
a = c(1, 2, 3)
is.numeric(a)
```

```
[1] TRUE
```

```
is.factor(a)
```

```
[1] FALSE
```

2.2.3 資源

- 資料結構

2.3 介紹

在 R 語言中, 型態不須經過宣告 (declared)。一個變數的型態經由 assignment 的過程決定, 即 <-右邊的 R-Objects。也就是在指派變數值的時候, 同時決定了型態。基本的 R-object 有 —

- Vectors
- Lists
- Matrices
- Arrays
- Factors
- Data Frames

最簡單的是 vector 物件,atomic vector 有 6 種 data types(有時也叫做 6 個 classes)

Data Type	Example
Logical	TRUE, FALSE
Numeric	1.3, 5, 99
Integer	3L, 24L, 0L
Complex	5 + 4i
Character	‘b’ , “good”, “TRUE”, ‘23.4’
Raw	“Hello” is stored as 48 65 6c 6c 6f

```
v <- TRUE
print(class(v))
```

```
[1] "logical"
```

```
v <- 23.5
print(class(v))
```

```
[1] "numeric"
```

```
v <- 2L
print(class(v))
```

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

```
v <- 2+5i
print(class(v))
```

```
[1] "complex"
```

```
v <- "TRUE"
print(class(v))
```

```
[1] "character"
```

```
v <- charToRaw("Hello")
print(class(v))
```

```
[1] "raw"
```

2.4 實數的比較

```
x <- seq(0, 1, by = 0.2)
y <- seq(0, 1, by = 0.2)
y[4]
```

```
[1] 0.6
```

```
x[3]
```

```
[1] 0.4
```

```
1 - x[3]
```

```
[1] 0.6
```

```
y[4] == 1 - x[3]
```

```
[1] FALSE
```

```
y[4] > 1 - x[3]
```

```
[1] TRUE
```

```
## note:
all.equal(y[4], 1 - x[3])
```

```
[1] TRUE
```

```
## Q: what is the result of : 1-0.4 ==0.6
```

```
0.1+0.2 == 0.3
```

```
[1] FALSE
```

```
all.equal(0.1+0.2,0.3)
```

```
[1] TRUE
```

2.5 字串

參考

2.5.1 建立字串

可以是雙引號中”” 或單引號中”。字串中如果有雙引號, 或單引號，則如下表示:
“ ‘這個’ 來自’ 那個’ ”

```
a <- "hello"
a
```

```
[1] "hello"
```

```
typeof(a)
```

```
[1] "character"
```

利用函數:character() 這個函數的參數, 為整數, 建立一個 list, 裡面的元素都是空字串

```
# 變數 ex 初始化為 character vector, 參看後面的討論
(ex <- character(0))
```

```
character(0)
```

```
length(ex)
```

```
[1] 0
```

```r
class(ex)
```

```
[1] "character"
```

```r
# 如果剛剛沒有設定 ex <- character(0), 這裡會發生錯誤
(ex[1] <- "first")
```

```
[1] "first"
```

```r
# check its length again
length(ex)
```

```
[1] 1
```

索引可以用跳的:

```r
(ex[4] <- "fourth")
```

```
[1] "fourth"
```

```r
length(ex)
```

```
[1] 4
```

```r
typeof(ex)
```

```
[1] "character"
```

```r
ex
```

```
[1] "first"  NA       NA       "fourth"
```

跳過的索引, 內容自動為 `NA`.

2.5.2 空字串

引號內連空白都沒有的字串: (比較上面利用 character(5) 可以建立 5 個元素為空字串的 vector。)

```
# empty string
empty_str <- ""
empty_str
```

```
[1] ""
```

```
# class
class(empty_str)
```

```
[1] "character"
```

2.5.2.1 討論 character(0)

前面說 character(2), 可以傳回長度 2, 每個元素都是空白字串”” 的向量, 那麼 character(0) 是甚麼? 除了前面提到的變數初始化為向量 (也許可以說是向量宣告) 例如, 整數也是這樣

```
zz<-integer(0)
zz[4]=6
zz
```

```
[1] NA NA NA  6
```

這裡對 character(0) 做一些測試:

```
ex1<-character(0)
ex2<-""

typeof(ex1)
```

```
[1] "character"
```

```
typeof(ex2)
```

```
[1] "character"
```

```
str(ex1)
```

```
 chr(0)
```

```
str(ex2)
```

```
 chr ""
```

```
class(ex1)
```

```
[1] "character"
```

```
class(ex2)
```

```
[1] "character"
```

```
is.list(ex1)
```

```
[1] FALSE
```

```
is.list(ex2)
```

```
[1] FALSE
```

surprise: 一個字元也是向量。

```
is.vector(ex1)
```

```
[1] TRUE
```

```
is.vector(ex2)
```

```
[1] TRUE
```

```
length(ex1)
```

```
[1] 0
```

```
length(ex2)
```

```
[1] 1
```

最後, 這兩個是不是相等

```
ex1==ex2
```

```
logical(0)
```

2.6 型態操作 is family

is.numeric(), is.integer(), and is.double() ## 型態轉換 as family

```
a<-c(TRUE,FALSE)
as.numeric(a)
```

```
[1] 1 0
```

```
an<-as.logical(a)
an
```

```
[1]  TRUE FALSE
```

2.7 vector

利用 `c` 函數，可以使用 vector 存放一個以上的數字。

```
a = c(1, 2, 3, 4, 5)
a1 = 1:5
```

有關 list 的運算: 加減乘除等等

```
a = c(1, 2, 3, 4, 5)
a+1
```

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

```
mean(a)
```

```
[1] 3
```

```
var(a)
```

```
[1] 2.5
```

```
summary(a)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1       2       3       3       4       5 
```

list 的元素, 利用中括號

```
a = 2:6
a[1]
```

```
[1] 2
```

```
a[2]
```

```
[1] 3
```

```
a[0]
```

```
integer(0)
```

```
a[6]
```

```
[1] NA
```

在 R 中, 的最小的索引值為 1base. 如果給索引為 0, 可以知道資料是否被排序。超出索引範圍得到”NA。

```
a=2:6
x = a[6]
```

如何判斷是否 NA

```
x == NA
```

```
[1] NA
```

上面的比較沒有意義, 和 NA 的任何運算都是 NA

```
r = x == NA
r
```

```
[1] NA
```

結論: 任何變數和 NA 運算, 結果還是 NA
另一種方法

```
print(x == NA)
```

```
[1] NA
```

如何判斷 NA ? `is.na()`

```
    is.na(a[6])
```

```
[1] TRUE
```

初始化向量, 可以利用 a<-10 或指定 numeric(double) 型態

```
a <- numeric(10)
a
```

```
 [1] 0 0 0 0 0 0 0 0 0 0
```

如果想要知到變數的資料型別, 利用函數 typeof()

typeof() 函數回傳的結果是 “字串”

```
typeof(a) # 結果是"double"
```

```
[1] "double"
```

```
s = typeof(a)
s
```

```
[1] "double"
```

```
typeof(s) # 結果是 "character"
```

```
[1] "character"
```

2.7.1 練習範例

Q1. a,a1,a2 屬於甚麼型態

```
a = 1:4
a1 = c(1, 2, 3, 4)
a2 = numeric(4)
```

A1

Q2:a3 的長度是甚麼?2, 或 6

```
a1<-c(1,2,3)
a2<-c(2,3,4)
a3<-c(a1,a2)
```

HINT: a1 a2 a3;length(a3)

2.8 字串和 vector

在 EXCEL 中，vector 一般指的是只放元素為數字的陣列 (array); 而陣列是可以存數字和文字的區域。table 是有欄位的陣列。但是在 R 語言中,vector 只是元素型態相同即可。

```
a <- "hello"
a
```

```
[1] "hello"
```

```
typeof(a)
```

```
[1] "character"
```

```
b <- c("hello", "there")
b
```

```
[1] "hello" "there"
```

```
b[1]
```

```
[1] "hello"
```

```
typeof(b)
```

```
[1] "character"
```

```
(a = character(5)) # 產生 5 個空白字串
```

```
[1] "" "" "" "" ""
```

```
(b = letters[1:4]) # 注意,letters 不是函數
```

```
[1] "a" "b" "c" "d"
```

```
letters
```

```
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p"
[17] "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"
```

因為 c 函數的運算結果為 vector, 因此下例中, 其元素都是字串

```
 (a<-c("d",4,TRUE))
```

```
[1] "d"    "4"    "TRUE"
```

問題: 怎樣知道 r 是空集合?

```
y <- letters[1:3]
z <- letters[4:6]
r<-intersect(y,z)
r
```

```
character(0)
```

```
is.na(r)
```

```
logical(0)
```

另外, 當 vector 有多個字串, 而使用索引 0 的時候, 也會出現 `character(0)`, 例如:

```
string <- c('sun', 'sky', 'clouds')
string[0]
```

```
character(0)
```

2.8.1 vector 相關的運算

連續數字可以利用操作元:, 例如:

x <- 1:7; x y <- 2:-2; y

比較複雜的序列可以利用函數 seq() , 例如

```
seq(1, 3, by=0.2)          # specify step size
```

```
 [1] 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
```

```
 seq(1, 5, length.out=4)    # specify length of the vector
```

```
[1] 1.00 2.33 3.67 5.00
```

2.8.2 如何存取 vector 中的元素

元素索引可以利用 logical, integer or character vector.

如果利用整數索引, 則從 1 開始. 但是, 如果索引值給的是負數 (例如-2), 則除了 2 號元素以外, 都會被傳回。但是不能同時有正負。同時, 浮點數會被 truncated。

```
> x
[1]  0  2  4  6  8 10
> x[3]             # access 3rd element
[1] 4
> x[c(2, 4)]       # access 2nd and 4th element
[1] 2 6
> x[-1]            # access all but 1st element
[1]  2  4  6  8 10
> x[c(2, -4)]      # 不能混合正負
Error in x[c(2, -4)] : only 0's may be mixed with negative subscripts
> x[c(2.4, 3.54)]    # real numbers are truncated to integers
[1] 2 4
```

2.8.3 邏輯做為索引

說是索引有點誤導, 可以認為是元素篩選。例如

```
> x[c(TRUE, FALSE, FALSE, TRUE)]
[1] -3  3
> x[x < 0]  # filtering vectors based on conditions
[1] -3 -1
> x[x > 0]
[1] 3
```

In the above example, the expression x>0 will yield a logical vector (FALSE, FALSE, FALSE, TRUE) which is then used for indexing.

2.8.4 利用字串 (character vector) 作為索引

每個元素可以有名稱:

```
> x <- c("first"=3, "second"=0, "third"=9)
> names(x)
[1] "first"  "second" "third"
> x["second"]
second
0
> x[c("first", "third")]
first third
3     9
```

```
a<-c(x=1:2,y=3:4)
a["x1"] # 不是 a[x1]
```

```
x1
 1
```

和 [] 的差別:
原來是甚麼 (例如,list 或 vector), 只是返回子集合 (仍然是 list 或 vector), 但是 [] 則是返回內容.

```
x <- c(a = 1, b = 2, c = 3)
x["a"]
```

```
a
1
```

```
x[["a"]]
```

```
[1] 1
```

```
x[1]
```

```
a
1
```

```
x[[1]]
```

```
[1] 1
```

和 list 的區別是 1. $ 1. 不必有""

```
a1<-list(x=1:2,y=3:4)
a1$x
```

```
[1] 1 2
```

2.8.5 How to modify a vector in R?

We can modify a vector using the assignment operator.

We can use the techniques discussed above to access specific elements and modify them.

If we want to truncate the elements, we can use reassignments.

```
> x
[1] -3 -2 -1  0  1  2
> x[2] <- 0; x        # modify 2nd element
[1] -3  0 -1  0  1  2
> x[x<0] <- 5; x   # modify elements less than 0
[1] 5 0 5 0 1 2
> x <- x[1:4]; x      # truncate x to first 4 elements
[1] 5 0 5 0
```

2.8.6 How to delete a Vector?

We can delete a vector by simply assigning a NULL to it.

```
> x
[1] -3 -2 -1  0  1  2
> x <- NULL
> x
NULL
> x[4]
NULL
```

2.9 List

Lists 是一個 R 物件。其元素內容可以是不同型態, 例如 numbers, strings, vectors,matrix 或者是另一個 list, 甚至是函數。。

2.9.1 建立 List

函數 `list()` 可以建立 list.

包含多種 type 的 List: strings, numbers, vectors and a logical values.

```
list_data <- list("Red", "Green", c(21,32,11), TRUE, 51.23, 119.1)
print(list_data)
```

```
[[1]]
[1] "Red"

[[2]]
[1] "Green"

[[3]]
[1] 21 32 11

[[4]]
[1] TRUE

[[5]]
[1] 51.2

[[6]]
[1] 119
```

### Naming List Elements

The list elements can be given names and they can be accessed using these names.

```
#Create a list containing a vector, a matrix and a list.
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2),
   list("green",12.3))

#Give names to the elements in the list.
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")

#Show the list.
print(list_data)
```

```
$`1st Quarter`
[1] "Jan" "Feb" "Mar"

$A_Matrix
     [,1] [,2] [,3]
[1,]    3    5   -2
[2,]    9    1    8

$`A Inner list`
$`A Inner list`[[1]]
[1] "green"

$`A Inner list`[[2]]
[1] 12.3
```

這裡討論一下為甚麼在 R 語言堅持 <-而不是 = 的一個原因

將列名的 vector 和 list , 白話文意思就是將每個元素命名。

```
x=c(1,2,3)
names(x)<-c('a','b','c')
x
```

```
a b c
1 2 3
```

一般的程式語言, 不會用到這種文法, 例如, 在其他程式語言中, 直覺上應該是有一個 names(), 然後用法如下:

```
names(x,c('a','b','c') )
```

但如果是 matrix, 被命名的就是欄位? 答案是:**NO**。而是紀錄名稱? 答案也是 **NO**. 而是給每個元素命名。

```
x<-matrix(runif(9),nrow=3)
names(x)<-c('a','b','c')
x
```

```
       [,1]   [,2]  [,3]
[1,] 0.0808 0.1572 0.498
[2,] 0.8343 0.0074 0.290
[3,] 0.6008 0.4664 0.733
attr(,"names")
[1] "a" "b" "c" NA  NA  NA  NA  NA  NA
```

```
rownames(x)<-c('x','y','z')
print(x,row.names=T)
```

```
    [,1]   [,2]  [,3]
x 0.0808 0.1572 0.498
y 0.8343 0.0074 0.290
z 0.6008 0.4664 0.733
attr(,"names")
[1] "a" "b" "c" NA  NA  NA  NA  NA  NA
```

問: 在上面的例子中 x 是一個矩陣,x[“a”] 會傳回甚麼?

但是在 data.frame 的結構中,names()<- 給定的是欄位名稱

```
xd<-data.frame(x)
names(xd)<-c('a','b','c')
xd
```

```
       a      b     c
x 0.0808 0.1572 0.498
y 0.8343 0.0074 0.290
z 0.6008 0.4664 0.733
```

2.9.2 存取 List 元素

```
# 3 個元素, 分別是 list, matrix, list
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2),
   list("green",12.3))

# 給名稱
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")
```

```
# Access the first element of the list.
print(list_data[1])
```

```
$`1st Quarter`
[1] "Jan" "Feb" "Mar"
```

```
# Access the thrid element. As it is also a list, all its elements will be printed.
print(list_data[3])
```

```
$`A Inner list`
$`A Inner list`[[1]]
[1] "green"

$`A Inner list`[[2]]
[1] 12.3
```

```
# Access the list element using the name of the element.
print(list_data$A_Matrix)
```

```
     [,1] [,2] [,3]
[1,]    3    5   -2
[2,]    9    1    8
```

2.9.3 Manipulating List Elements

We can add, delete and update list elements as shown below. We can add and delete elements only at the end of a list. But we can update any element.

```
# Create a list containing a vector, a matrix and a list.
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2),
   list("green",12.3))

# Give names to the elements in the list.
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")

# Add element at the end of the list.
list_data[4] <- "New element"
print(list_data[4])
```

```
[[1]]
[1] "New element"
```

```
# Remove the last element.
list_data[4] <- NULL

# Print the 4th Element.
print(list_data[4])
```

```
$<NA>
NULL
```

```
# Update the 3rd Element.
list_data[3] <- "updated element"
print(list_data[3])
```

```
$`A Inner list`
[1] "updated element"
```

2.9.4 Merging Lists

You can merge many lists into one list by placing all the lists inside one list() function.

```
# Create two lists.
list1 <- list(1,2,3)
list2 <- list("Sun","Mon","Tue")

# Merge the two lists.
merged.list <- c(list1,list2)

# Print the merged list.
print(merged.list)
```

```
[[1]]
[1] 1

[[2]]
[1] 2

[[3]]
[1] 3

[[4]]
[1] "Sun"
```

```
[[5]]
[1] "Mon"

[[6]]
[1] "Tue"
```

我們說函數 c() 是一個處理向量的函數, 但是這裡的 list 裡面有多個型態, 所以為甚麼用 c()? 因為, 對 c() 而言, 兩個元素都是 list 。而在 R 語言中, 向量的意思, 只要元素同型態就行。

2.9.5 List 轉 Vector

函數 unlist()

```
# Create lists.
list1 <- list(1:5)
print(list1)
```

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

```
list2 <-list(10:14)
print(list2)
```

```
[[1]]
[1] 10 11 12 13 14
```

```
# Convert the lists to vectors.
v1 <- unlist(list1)
v2 <- unlist(list2)

print(v1)
```

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

```
print(v2)
```

```
[1] 10 11 12 13 14
```

```
# Now add the vectors
result <- v1+v2
print(result)
```

```
[1] 11 13 15 17 19
```

2.9.6 比較 vector, list 中, 字串的問題

1. Q length() 函數可以知道 vector,list 的長度, 但是為甚麼 length(“hello”) 的長度是 1?

```
a<-c("hello","r")
length(a)
```

```
[1] 2
```

```
length(a[1])
```

```
[1] 1
```

```
length(a[[1]])
```

```
[1] 1
```

```
nchar(a[1])
```

```
[1] 5
```

```
length("hello")
```

```
[1] 1
```

1. Q

```
a<-list("hello","r")
length(a)
```

```
[1] 2
```

```
length(a[1])
```

```
[1] 1
```

```
length(a[[1]])
```

```
[1] 1
```

上面兩個答案的問題, 和 R 怎樣建立字串有關我猜是 “hello” 的記憶體結構是 list(ptr) ,ptr->“hello” 也就是説紀錄在 list 中, 當我問説 length(“hello”) 的時候,list 的長度都是 1。
總之, 字串長度函數 `nchar()`。

2.10 data set build in

2.10.1 List of pre-loaded data

```
data()
```

Loading a built-in R data

```
data(mtcars)
```

```
nrow(mtcars)
```

```
[1] 32
```

```
ncol(mtcars)
```

```
[1] 11
```

2.11 vector 函數範例

2.11.1 cbind,rbind

```
x<-runif(5)
y<-runif(6)
cbind(x,y)
```

```
Warning in cbind(x, y): number of rows of result is not a multiple of
vector length (arg 1)
```

```
          x     y
[1,] 0.0808 0.466
[2,] 0.8343 0.498
[3,] 0.6008 0.290
[4,] 0.1572 0.733
[5,] 0.0074 0.773
[6,] 0.0808 0.875
```

2.11.2 函數 diff

Arguments
* x: a numeric vector or matrix containing the values to be differenced.
* lag: an integer indicating which lag to use.
* differences: an integer indicating the order of the difference.
例如 diff(x,lag=d,differences=n) x[(1+d):n] - x[1:(n-d)]

2.12 練習: 刪除向量中的一個元素

```
x<-c(1,2,3)
```

如何刪除 x 中的第 2 個元素?

hint:
數字向量元素的刪除，不能像是 list[2]<-NULL ，請問你的解法? 並説明

物件元素的刪除，簡單一點的像是 list，

```
x<-list(1,2,3)
x[2]<-NULL
x
```

但是向量元素的刪除沒有捷徑

```
x<-c(1,2,3)
x[2]<-NULL
```

```
Error in x[2] <- NULL:
  置換的長度為零
```

因此，需要變通的辦法
方法 1:

```
x<-c(1,2,3)
t<-as.list(x) # note: 使用 list(x) 而不是 as.list(x) 會有什麼不同
t[2]<-NULL
x<-unlist(t)
x
```

```
[1] 1 3
```

方法 2:

```
x<-c(1,2,3)
y<-x[-2]
x<-y
```

練習: 改為函數

```
delElement<-function(x,idx){
  if (is.vector(x)){
    t<-as.list(x)
    t[idx]<-NULL
    t<-unlist(t)
    return(t)
  } else {
    stop('must be vector')
  }

}
x<-c(1,2,3)
x<-delElement(x,2)
x
```

```
[1] 1 3
```

2.12.1 Adding Attributes to Vectors

The attributes that you can add to vectors includes names and comments. If we continue with our vector v1 we can see that the vector currently has no attributes:

```
v1 <- 8:17
attributes(v1)
```

```
NULL
```

We can add names to vectors using two approaches. The first uses names() to assign names to each element of the vector. The second approach is to assign names when creating the vector.

```
# assigning names to a pre-existing vector
names(v1) <- letters[1:length(v1)]
v1
```

```
 a  b  c  d  e  f  g  h  i  j 
 8  9 10 11 12 13 14 15 16 17 
```

```
attributes(v1)
```

```
$names
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

那，什麼是 `names(v1)`

```
names(v1)
```

```
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

也就是說 names() 函數返回每個元素的名子，而 `attributes()` 返回的是有什麼屬性，也是函數名稱字面上的意思。但是再仔細看看，可以發現指派 attribute 給

```
attributes(v1)<-list(x='',y='',z='')
v1
```

```
 [1]  8  9 10 11 12 13 14 15 16 17
attr(,"x")
[1] ""
attr(,"y")
[1] ""
attr(,"z")
[1] ""
```

測試是否可以自行加入 attributes，然後讓 `names()` 可以運作，如下：

```
attributes(v1)<-list(names=letters[1:10])
v1
```

```
 a  b  c  d  e  f  g  h  i  j 
 8  9 10 11 12 13 14 15 16 17 
```

```
attributes(v1)
```

```
$names
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

```
names(v1)
```

```
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

上面這個程式暗示了什麼? > 函數 names() 只是 attributes()<-的一個特例，我們總是可以自行加入一個屬性，名稱為 names 讓函數 names() 可以把每個元素的名稱列印出來。

後面這段程式碼只是對照一下之前每個元素加入名稱的作法，本身沒意義：

```
# adding names when creating vectors
v2 <- c(name1 = 1, name2 = 2, name3 = 3)
v2
```

```
name1 name2 name3
    1     2     3
```

```
attributes(v2)
```

```
$names
[1] "name1" "name2" "name3"
```

也可以加入註解：

```
comment(v1) <- "This is a comment on a vector"
v1
```

```
 a  b  c  d  e  f  g  h  i  j
 8  9 10 11 12 13 14 15 16 17
```

```
attributes(v1)
```

```
$names
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"

$comment
[1] "This is a comment on a vector"
```

2.13 matrix

矩陣的建立有多種方式, 其中一種是利用向量轉填矩陣，填入的方式預設是以 coumn 為主要方向。

```
matrix(c(1,2,3,4,5,6),2,3)
```

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

問題: 怎樣產生
1 2 3
4 5 6
的矩陣

```
diag(1, nrow = 5)
```

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

```
matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), nrow = 3, byrow = TRUE, dimnames = list(c("r1", "
```

```
   c1 c2 c3
r1  1  2  3
r2  4  5  6
r3  7  8  9
```

```
m1 <- matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), ncol = 3)
rownames(m1) <- c("r1", "r2", "r3")
colnames(m1) <- c("c1", "c2", "c3")

m1
```

```
   c1 c2 c3
r1  1  4  7
r2  2  5  8
r3  3  6  9
```

```
 B = matrix(
    c(2, 4, 3, 1, 5, 7),
    nrow=3,
    ncol=2)
B
```

```
     [,1] [,2]
[1,]    2    1
[2,]    4    5
[3,]    3    7
```

2.13.1 Transpose

```
t(B)
```

```
     [,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7
```

2.13.2 合併矩陣

利用函數 `cbind()` 可以合併同樣橫行數目的兩個矩陣, 例如這裡 C 也是 3 個橫行:

```
C = matrix(
   c(7, 4, 2),
   nrow=3,
   ncol=1)
C
```

```
     [,1]
[1,]    7
[2,]    4
[3,]    2
```

和 B 合併

```
cbind(B, C)
```

```
     [,1] [,2] [,3]
[1,]    2    1    7
[2,]    4    5    4
[3,]    3    7    2
```

同樣的,`rbind()` 也可以合併直行數目相同的兩個矩陣:

```
 D = matrix(
   c(6, 2),
   nrow=1,
   ncol=2)
D
```

```
     [,1] [,2]
[1,]    6    2
```

```
rbind(B, D)
```

```
     [,1] [,2]
[1,]    2    1
[2,]    4    5
[3,]    3    7
[4,]    6    2
```

將矩陣解構回向量: 可以利用 `c()` 會把矩陣所有的直行都合併成一個直行.

```
c(B)
```

```
[1] 2 4 3 1 5 7
```

unlist ?

練習: 利用上面的 m1 回答下面的問題:
m1[1, 2]

m1[1:2, 2:3]

m1[1,] m1[,2] m1[1:2,] m1[, 2:3]

m1[-1,] m1[,-2]

m1[c("r1", "r3"), c("c1", "c3")]

m1[1] m1[9] m1[3:7]

m1 > 3

m1[m1 > 3]

m1 + m1 m1 - 2*m1 m1* m1 m1 / m1 m1 ^ 2 m1 %*% m1

t(m1)

2.14 字串函數

2.14.0.1 asic character string functions provided by R:

- nchar: string length

- paste: concatenate strings
- substr: substring
- toupper: convert entire string to uppercase
- tolower: convert entire string to lowercase
- chartr: character map replacement (like “tr”)
- strtrim： trunates string

nchar, substr, toupper, tolower will accept string vectors as arguments and return vector results.

strtrim accepts both a vector of strings and a vector of truncation positions.

- `\'`: 等同於 `"'"` .

- `\"`: 等同於 `'"'`.

- `\n`: newline.

- `\r`: carriage return.

- `\t`: tab character.

> Note: `cat()` and `print()` 處理逃逸字元的方式不一樣如果要在螢幕上解讀上面的字串, 需要的是 `cat()`.

```r
print("a\nb")
```

```
[1] "a\nb"
```

```r
cat("a\nb")
```

```
a
b
```

問題: 1. 解釋 paste0(c(1,2),c(3,4),collapse=”) 2. 怎樣得到 “1234”

2.14.1 應用範例 macro

version 1

```
x<- "${1} is good"
gsub("\\$\\{1\\}","dog",x)
```

```
[1] "dog is good"
```

** version 2 to function**

```
mstr<-function (src,mto)
{
  return (gsub("\\$\\{1\\}",mto,src))

}
x<- "${1} is good"
mstr(x,"dog")
```

```
[1] "dog is good"
```

2.14.2 paste

```
paste0("x","y","z")
```

```
[1] "xyz"
```

2.14.3 misc

.

The get() and assign() functions allow you to reference objects by character strings, and as.name() will convert a string into a reference (you would then probably need to use substitute() to create an expression and eval()) to evaluate it.

eg

```
assign("a",get("b"))
```

does a<-b

Often there is a better way: the objects can be stored in a list, making it possible to use character strings directly to reference them

If obj<-list(a=1,b=2)

then obj[[“a”]] is a reference to the “a” element and you can do eg

obj[[“a”]]<-otherobj[[“b”]]

2.15 Factors

R 也能把資料存為 factor(store data is as a factor)。在大部分實驗中, 某些解釋變數經常有不同程度的測試。大意如下:
>… includes trials for different levels of some explanatory variable. >The different levels are also called factors.

2.15.1 探索一下 factors

```
fert = c(10,20,20,50,10,20,10,50,20)
fert = factor(fert,levels=c(10,20,50),ordered=TRUE)
```

levels() 可以顯示 factor 的標籤

```
levels(fert)
```

```
[1] "10" "20" "50"
```

fert 的內容, 如果忽略標籤本身, 其實已經是數字。

```
fert
```

```
[1] 10 20 20 50 10 20 10 50 20
Levels: 10 < 20 < 50
```

計算平均

```
mean(fert)
```

```
Warning in mean.default(fert): argument is not numeric or logical:
returning NA
```

```
[1] NA
```

```r
mean(as.numeric(levels(fert)[fert]))
```

```
[1] 23.3
```

解釋:
levels(fert) 產生一個「字串向量」, 內容為 10, 20, 50, 如下可證:

```r
str(levels(fert))
```

```
 chr [1:3] "10" "20" "50"
```

然後如果把 fert 想程式數字內容, 例如 (1,2,1) 則, 這是索引。就像這樣
c("a","b","c")[c(1,1,2,3,2)] ==> a,a,b,c,b 。

上面的問題, 可以簡化為: 找回原來還沒被 factor 的向量。

```r
x<-c("a","b","c")
x<-factor(x) # 現在 x 是一個 factor

x<-levels(x)[x] # 現在又被轉回原來的 x
```

2.15.2 探索一下 factor

```r
a <- factor(c("A", "A", "B", "A", "B", "B", "C", "A", "C"))
```

預期答案是 factor,ok。

```r
class(a)
```

```
[1] "factor"
```

因為 a 已經被編碼, 預期是 numeric, 這裡答案直接給定 integer:

```r
typeof(a)
```

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

a 是一個 factor, 所以 a 代表 a 的一個子集合, 因此結果看起來也是類似原來 a 的型態:

```
a[1]
```

```
[1] A
Levels: A B C
```

驗證是不是和上面的表示方式一樣:

```
a
```

```
[1] A A B A B B C A C
Levels: A B C
```

所以，a,a 都是類型為 factor 的變數。

按照之間 `[]` 和 `[[]]` 的規則，這裡會出現什麼?

```
a[[1]]
```

```
[1] A
Levels: A B C
```

出現的結果和 a[1] 一樣，這說明了，a 的結構不是 list, 利用 str() 函數看一下確認:

```
str(a)
```

```
 Factor w/ 3 levels "A","B","C": 1 1 2 1 2 2 3 1 3
```

總之，a[1] a[[1]] 都是 Factor 沒有 [和 [[的分別。這很好記，因為這個 factor 也是 r 的內建資料型態。
其他驗證如下，自行推敲:

又有新的問題了，到底 table(a),summary(a) 的資料結構是什麼?

```
str(table(a))
```

```
 'table' int [1:3(1d)] 4 3 2
 - attr(*, "dimnames")=List of 1
  ..$ a: chr [1:3] "A" "B" "C"
```

```
str(summary(a))
```

```
 Named int [1:3] 4 3 2
 - attr(*, "names")= chr [1:3] "A" "B" "C"
```

```r
class(table(a))
```

```
[1] "table"
```

```r
class(summary(a))
```

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

又來一個讓人無言的地方，輸出樣子看起來很像，但是結構一個是類別 `table` 一個是「整數向量」，然後每個元素都是因子的發生次數，帶有因子名字，這裡是 “A”,“B”,“C”。

2.15.3 factor 和函數 read.csv()

read.csv() 通常產生一個 data.frame 的類別。但是在讀檔的時候，如果遇到字串欄位，會將字串欄位預設轉為 factor 欄位。

觀察 carbon dioxide 對樹木的生長速率 trees91.csv 1^:

```r
tree = read.csv("./resources/trees91.csv", header = TRUE, sep = ",");
attributes(tree)
```

```
$names
 [1] "C"       "N"       "CHBR"    "REP"     "LFBM"    "STBM"    "RTBM"
 [8] "LFNCC"   "STNCC"   "RTNCC"   "LFBCC"   "STBCC"   "RTBCC"   "LFCACC"
[15] "STCACC"  "RTCACC"  "LFKCC"   "STKCC"   "RTKCC"   "LFMGCC"  "STMGCC"
[22] "RTMGCC"  "LFPCC"   "STPCC"   "RTPCC"   "LFSCC"   "STSCC"   "RTSCC"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
[22] 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
[43] 43 44 45 46 47 48 49 50 51 52 53 54
```

```r
names(tree)
```

```
 [1] "C"       "N"       "CHBR"    "REP"     "LFBM"    "STBM"    "RTBM"
 [8] "LFNCC"   "STNCC"   "RTNCC"   "LFBCC"   "STBCC"   "RTBCC"   "LFCACC"
[15] "STCACC"  "RTCACC"  "LFKCC"   "STKCC"   "RTKCC"   "LFMGCC"  "STMGCC"
[22] "RTMGCC"  "LFPCC"   "STPCC"   "RTPCC"   "LFSCC"   "STSCC"   "RTSCC"
```

A description of the data file is located at http://cdiac.ornl.gov/ftp/ndp061a/ndp061a.txt .

2.15.4 factor 和函數 summary()

```r
summary(tree$CHBR)
```

```
 A1  A2  A3  A4  A5  A6  A7  B1  B2  B3  B4  B5  B6  B7  C1  C2  C3 
  3   1   1   3   1   3   1   1   3   3   3   3   3   3   1   3   1 
 C4  C5  C6  C7 CL6 CL7  D1  D2  D3  D4  D5  D6  D7 
  3   1   1   1   1   1   1   1   3   1   1   1   1 
```

在 CHBR 這個欄位中, 因為不全都是數字, 因此 R 自動假定這是一個 factor。因此針對這個欄位 `summary()` 函數不會列出 5 個統計量, 而是列出次數表。因為，一旦向量轉為一組 factors, 5 個基本統計量不再有意義。

但有些欄位例如 C, 也是一個 factor。但是,R 認定為數字, 這時必須手動處理。
以下將 tree$C 轉為 factor:

```r
tree$C
```

```
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3
[33] 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4
```

```r
summary(tree$C)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    2.00    2.00    2.52    3.00    4.00 
```

```r
tree$C <- factor(tree$C)
tree$C
```

```
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3
[33] 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
```

```r
summary(tree$C)
```

```
 1  2  3  4 
 8 23 10 13 
```

```r
levels(tree$C)
```

```
[1] "1" "2" "3" "4"
```

node: `summary()` 函數，參考 table 一節。

2.16 Data Frames

2.16.1 簡單介紹一下手動建立 data.frame

3 個欄位，分別是 a,b,f

```
a <- c(1, 2, 3, 4)
b <- c(2, 4, 6, 8)
levels <- factor(c("A", "B", "A", "B"))
bubba <- data.frame(first = a,
                      second = b,
                      f = levels)
bubba
```

```
  first second f
1     1      2 A
2     2      4 B
3     3      6 A
4     4      8 B
```

```
summary(bubba)
```

```
     first          second      f    
 Min.   :1.00   Min.   :2.0   A:2  
 1st Qu.:1.75   1st Qu.:3.5   B:2  
 Median :2.50   Median :5.0        
 Mean   :2.50   Mean   :5.0        
 3rd Qu.:3.25   3rd Qu.:6.5        
 Max.   :4.00   Max.   :8.0        
```

```
bubba$first
```

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

```
bubba$second
```

```
[1] 2 4 6 8
```

```
bubba$f
```

```
[1] A B A B
Levels: A B
```

2.16.2 data frame 的常用函數:

- head() - shows first 6 rows
- tail() - shows last 6 rows
- dim() - returns the dimensions of data frame (i.e. number of rows and number of columns)
- nrow() - number of rows
- ncol() - number of columns
- str() - structure of data frame - name, type and preview of data in each column
- names() - shows the names attribute for a data frame, which gives the column names.
- sapply(dataframe, class) - shows the class of each column in the data frame

回到之前利用 read.csv 讀入的 data.frame：tree

```
# 順便看看結構
typeof(tree)
```

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

```
class(tree)
```

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

```
is.data.frame(tree)
```

```
[1] TRUE
```

既然 tree 是一個 list, 那麼 `[]` 和 `[[]]` 的應用就和之前討論的一樣:

```
tree[1]
```

```
  C
1 1
2 1
3 1
4 1
5 1
6 1
7 1
8 1
```

```
9  2
10 2
11 2
12 2
13 2
14 2
15 2
16 2
17 2
18 2
19 2
20 2
21 2
22 2
23 2
24 2
25 2
26 2
27 2
28 2
29 2
30 2
31 2
32 3
33 3
34 3
35 3
36 3
37 3
38 3
39 3
40 3
41 3
42 4
43 4
44 4
45 4
46 4
47 4
48 4
49 4
50 4
51 4
52 4
53 4
54 4
```

```r
tree[[1]]
```

```
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3
[33] 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
```

```r
class(tree[1]) # data.frame
```

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

```r
class(tree[[1]]) # integer
```

```
[1] "factor"
```

```r
typeof(tree[1]) # list
```

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

```r
typeof(tree$C) # integer
```

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

```r
typeof(tree[[1]]) # integer
```

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

tree[1] 是一個 list
tree[[1]] 外圍的 tree 首先被解讀為 list 類別, 然後 [1] 傳到類別

> [[vs [
> 由 typeof,class 看起來
> [[抽取 list 中的元素 [只是分割 list, 中的 subset

2.17 Tables

除了 data frame 以外, 還有 table 可以用來組織資料。這裡只看怎樣建立 table, 分析看其他章節。

2.17.1 One Way Tables

table() 指令:
這個指令通常用來建立因子之間的關聯表 (contingency table)。參數通常是一個「因子向量」(factor vector)。
例如建立一個單因子關聯表（one way table）：

```
a <- factor(c("A", "A", "B", "A", "B", "B", "C", "A", "C"))
table_a <- table(a)
```

table_a:

```
table_a
```

```
a
A B C
4 3 2
```

attributes 可以看出 table_a 裡面有哪些帶名元素

```
attributes(table_a)
```

```
$dim
[1] 3

$dimnames
$dimnames$a
[1] "A" "B" "C"


$class
[1] "table"
```

```
summary(table_a)
```

```
Number of cases in table: 9
Number of factors: 1
```

對照 table(a)，直接把 factor 變數，代入 summary() 看一下，結果是次數分配表（如下）

```
table(a)
```

```
a
A B C 
4 3 2 
```

```
summary(a)
```

```
A B C 
4 3 2 
```

問題：如果我們知道 A 有 4, 個,B 有 3 個,C 有 2 個, 能不能直接建立 table?
1. 先建立 matrix
2. 再加入欄位名稱
3. 利用函數 as.table()

```
# step 1
occur <- matrix(c(4, 3, 2), ncol = 3, byrow = TRUE)
occur
```

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

```
#step 2
colnames(occur) <- c("A", "B", "C")
occur
```

```
     A B C
[1,] 4 3 2
```

```
#step 3
occur <- as.table(occur)
occur
```

```
  A B C
A 4 3 2
```

```
attributes(occur)
```

```
$dim
[1] 1 3
```

```
$dimnames
$dimnames[[1]]
[1] "A"

$dimnames[[2]]
[1] "A" "B" "C"


$class
[1] "table"
```

2.17.2 Two Way Tables

這個例子中有兩個問題: 第 1 個問題的答案有 “Never,” “Sometimes,” or “Always.” 第 2 個問題的答案有 “Yes,” “No,” or “Maybe.” 兩個問題分別以向量 a,b 存放 (The set of vectors “a,” and “b,” contain the response for each measurement.)

```
a <- c("Sometimes", "Sometimes", "Never", "Always", "Always", "Sometimes", "Sometimes"
b <- c("Maybe", "Maybe", "Yes", "Maybe", "Maybe", "No", "Yes", "No")
results <- table(a, b)
results
```

```
           b
a           Maybe No Yes
  Always        2  0   0
  Never         0  1   1
  Sometimes     2  1   1
```

在表格中, 可以看到同時回答 “Maybe” “Sometimes” 的個數有幾個。

這裡是另一個直接由我們知道的數據建立 table 的例子

```
sexsmoke <- matrix(c(70, 120, 65, 140), ncol = 2, byrow = TRUE)
rownames(sexsmoke) <- c("male", "female")
colnames(sexsmoke) <- c("smoke", "nosmoke")
sexsmoke <- as.table(sexsmoke)
sexsmoke
```

```
       smoke nosmoke
male      70     120
female    65     140
```

2.18 OOP

advance R ### Base objects vs OO objects 辨別變數是基礎型別或是物件型別，可以簡單利用函數：`is.object()`:

基礎物件 base object:

```
is.object(1:10)
```

```
[1] FALSE
```

[1] FALSE

物件導向物件 OO object

```
is.object(mtcars)
```

```
[1] TRUE
```

[1] TRUE

主要的區別在於基本物件沒有 class 這個 attribute

```
attr(1:10, "class") #  NULL
```

```
NULL
```

```
attr(mtcars, "class") # [1] "data.frame"
```

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

class() 這個函數, 不見得總是會和 attr() 的結果一致, 因為，對基本物件而言, 傳回的是後面討論, 而不是 NULL。

2.18.1 變數屬性查詢函數

names() dimnames() length() dim(): 各維度長度 class() : 變數類別 table()：各類資料計數 str()：變數的資料結構 (其實是一個 list)

透過 names() 函數，可取得各種 list 元素的名稱, 但是如果是 data.frame 則是顯示欄位名稱 islands 是內建資料

```
str(islands)
```

```
 Named num [1:48] 11506 5500 16988 2968 16 ...
 - attr(*, "names")= chr [1:48] "Africa" "Antarctica" "Asia" "Australia" ...
```

```
head(names(islands))
```

```
[1] "Africa"       "Antarctica"   "Asia"         "Australia"
[5] "Axel Heiberg" "Baffin"
```

方形資料（例如，矩陣，table，data.frame），可以透過 dimnames() 函數可顯示 data.frame 橫行和直行的名稱先橫行，然後直行。

```
dimnames(USArrests)
```

```
[[1]]
 [1] "Alabama"        "Alaska"         "Arizona"
 [4] "Arkansas"       "California"     "Colorado"
 [7] "Connecticut"    "Delaware"       "Florida"
[10] "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"
[16] "Kansas"         "Kentucky"       "Louisiana"
[19] "Maine"          "Maryland"       "Massachusetts"
[22] "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"
[28] "Nevada"         "New Hampshire"  "New Jersey"
[31] "New Mexico"     "New York"       "North Carolina"
[34] "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"
[40] "South Carolina" "South Dakota"   "Tennessee"
[43] "Texas"          "Utah"           "Vermont"
[46] "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"

[[2]]
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

方形資料，可以透過 length() 函數可顯示資料長度，但是如果是 data.frame 則顯示欄位數。

```
length(islands)
```

```
[1] 48
```

透過 dim() 顯示方形資料的「資料筆數」和「欄位數目」, 先「橫行」, 後「直行」

```
dim(USArrests)
```

```
[1] 50  4
```

使用 class() 函數可知道變數類別

```
class(pi)
```

```
[1] "numeric"
```

```
class(Sys.Date())
```

```
[1] "Date"
```

使用 table() 函數可知道向量中每個值出現幾次

```
table(iris$Species) ## 統計結果
```

```

    setosa versicolor  virginica 
        50         50         50 
```

2.19 operator %>%

%>% 不是 R 基礎套件而是定義再套件 magrittr (CRAN) 且常跟 dplyr (CRAN) 搭配。

意思是把左邊 (LHS) 當成右邊 (RHS) 的參數。
例如下面的例子: 資料框 iris 用來當 head() 的參數: 也就是說 iris %>% head() 相當於 head(iris).

library(magrittr) iris %>% head() Sepal.Length Sepal.Width Petal.Length Petal.Width Species 1 5.1 3.5 1.4 0.2 setosa 2 4.9 3.0 1.4 0.2 setosa 3 4.7 3.2 1.3 0.2 setosa 4 4.6 3.1 1.5 0.2 setosa 5 5.0 3.6 1.4 0.2 setosa 6 5.4 3.9 1.7 0.4 setosa

為甚麼需要這樣用, 下面是一個例子

iris %>% head() %>% summary() 類似的觀念, iris %>% head() %>% summary() 等同於 summary(head(iris)). 也就是說, 避免了使用巢狀呼叫。

2.20 apply family

2.20.1 apply

可以處理的對象包括矩陣、資料集、陣列 (二維、多維)，可以指定直行、橫行。

函數定義：

```
apply(X, MARGIN, FUN, ...)
```

參數列表：- X : 矩陣 matrix、資料集 dataframe 、陣列 array - MARGIN : 1 表示橫行，2 表示直行 - FUN : the function to be applied
- … : optional arguments to FUN

範例 1

```
x<-matrix(1,ncol=4,nrow=3);x
```

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

```
apply(x,1,sum)
```

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

範例：

有一個 list, 裡面有 2 個欄位，分別是 x1,x2, 想要分別算出每個欄位的最小和最大

```
x <- cbind(x1 = 3, x2 = c(1:6)); # x 是矩陣
f1<- function(col) {
   c(min(col),max(col))
 }

 apply(x,2,f1)
```

```
     x1 x2
[1,]  3  1
[2,]  3  6
```

練習：如果用 for loop? 練習：哪種方式比較快? hint: 利用 system.time(a function)

2.20.2 lapply 函數

大綱：lapply(),sapply(),vapply() 都參數都吃 list 或者 vector。
* lapply(), sapply() 差別是前者返回 list, 後者返回 matrix; * sapply(), vapply() 差別是，後者可以加入 row names。

lapply 的參數是 list、data.frame，然後返回和 X 長度同樣的 list 結構作為結果。

函數定義：

```
lapply(X, FUN, ...)
```

參數列表：

- X: list、vector、data.frame
- FUN: 自定義的調用函數
- ...: 更多參數，可選

比如，計算 list 中的每個 KEY 對應的 5 個 Tukey 分位數 (minimum, lower-hinge, median, upper-hinge, maximum) 。利用 `fivenum()`

```
x <- list(a = 1:6, b = rnorm(2,6,8));x
```

```
$a
[1] 1 2 3 4 5 6

$b
[1] -5.20  8.04
```

```
lapply(x,fivenum)
```

```
$a
[1] 1.0 2.0 3.5 5.0 6.0

$b
[1] -5.20 -5.20  1.42  8.04  8.04
```

如果傳入的參數是 data.frame(), 如下，對資料集的每個欄位求和。

```
lapply(data.frame(x), sum)
```

```
$a
[1] 21

$b
[1] 8.53
```

note: 如果傳入的參數是矩陣，則會對每個元素求函數值。

```
x <- matrix(rnorm(4,5,10),nrow=2);x
```

```
       [,1] [,2]
[1,] -19.37 11.2
[2,]   4.94 16.5
```

```
lapply(x, sum)
```

```
[[1]]
[1] -19.4

[[2]]
[1] 4.94

[[3]]
[1] 11.2

[[4]]
[1] 16.5
```

2.20.3 sapply 函數

sapply 函數包裹了一上面提到的 `lapply()`，返回值為向量，而不是 list 對象。sapply 增加了 2 個參數 simplify 和 USE.NAMES，

函數定義：

```
sapply(X, FUN, ..., simplify=TRUE, USE.NAMES = TRUE)
```

參數列表：

- X: list、matrix、data.frame
- FUN: 自定義的調用函數
- ...: 更多參數，可選
- simplify: 預設為 T, 返為值為向量；如果是 F 成 list。? 設定為 "array" 時，輸出結果按數組進行分組
- USE.NAMES: 如果 X 為字符串，TRUE 設置字符串為數據名，FALSE 不設置

note: 如果我們利用函數 `unlist` 配合 lapply 的返回值，也可以得到向量，例如：`unlist(lapply(1:3, function(x) x^2))`

```
x <- cbind(x1 = 3, x2 = c(1:6)); # x 是矩陣
sapply(x, sum)  # 結果類似 lapply()
```

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

```
# data.frame
sapply(data.frame(x), sum)
```

```
x1 x2
18 21
```

```
# 確認返回類別
class(lapply(x, sum))
```

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

```
class(sapply(x, sum))
```

```
[1] "numeric"
```

```
class(sapply(x, sum, simplify=FALSE, USE.NAMES=FALSE))
```

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

說明書提到：simplify=FALSE 和 USE.NAMES=FALSE，則 sapply 函數= lapply 函數。

練習：解釋下面的輸出結果。

```
x <- cbind(x1 = 3, x2 = c(1:6))
sapply(x, quantile)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
0%      3    3    3    3    3    3    1    2    3     4     5     6
25%     3    3    3    3    3    3    1    2    3     4     5     6
50%     3    3    3    3    3    3    1    2    3     4     5     6
75%     3    3    3    3    3    3    1    2    3     4     5     6
100%    3    3    3    3    3    3    1    2    3     4     5     6
```

```
sapply(2:4, seq)
```

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

[[2]]
[1] 1 2 3

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

hint: sapply 的調用函數 lapply()，而已經知道 lapply() 的參數是矩陣的時候，每個元素都會被傳到函數。因此欄位有 12 個；又，每個 quantile() 傳回 5 個元素，因此，橫行有 5 個。正確的作法是：

```
x <- list(x1 = 3, x2 = c(1:6))
sapply(x, quantile)
```

```
     x1   x2
0%    3 1.00
25%   3 2.25
50%   3 3.50
75%   3 4.75
100%  3 6.00
```

範例：USE.NAMES=TRUE

```
alist<-list(a=c(1,1,1),b=c(2,2,2),c=c(3,3,3))
f2<-function(x)
{
  mean(alist[[x]])
}
sapply(c("a","b","c"),f2,USE.NAMES = T)
```

```
a b c
1 2 3
```

對於 simplify 為 array 時，我們可以參考下面的例子，構建一個三維數組，其中二個維度為方陣。

```
a<-1:2
# 按數組分組
sapply(a,function(x) matrix(x,2,2), simplify='array')
```

```
, , 1

     [,1] [,2]
[1,]    1    1
[2,]    1    1

, , 2

     [,1] [,2]
[1,]    2    2
[2,]    2    2
```

```
# 默認情況，則自動合併分組
sapply(a,function(x) matrix(x,2,2))
```

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

2.20.4 比較

Function	Arguments	Objective	Input	Output
apply	apply(x, MARGIN, FUN)	Apply a function to the rows or columns or both	Data frame or matrix	vector, list, array
lapply	lapply(X, FUN)	Apply a function to all the elements of the input	List, vector or data frame	list
sapply	sappy(X FUN)	Apply a function to all the elements of the input	List, vector or data frame	vector or matrix

```
m<-matrix(1:4,4,3);m
```

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

```
sapply(1:3, function(x) mean(m[,x]))
```

```
[1] 2.5 2.5 2.5
```

```
m<-matrix(1:4,4,3);
lapply(1:3, function(x) mean(m[,x]))
```

```
[[1]]
[1] 2.5

[[2]]
[1] 2.5

[[3]]
[1] 2.5
```

2.20.5 vapply 函數

vapply 類似於 sapply，提供了 FUN.VALUE 參數，用來控制返回值的行名。

note: `sapply()` 配上 `rownames<-` 可以達到一樣的效果。

函數定義：

```
vapply(X, FUN, FUN.VALUE, ..., USE.NAMES = TRUE)
```

參數列表：

- X: 數組、矩陣、數據框
- FUN: 自定義的調用函數
- FUN.VALUE: 可以記為對每一紀錄給名稱 (e.g. row.names)
- ...: 更多參數，可選
- USE.NAMES: 如果 X 為字符串，TRUE 設置字符串為數據名，FALSE 不設置

```
x <- cbind(x1 = 3, x2 = c(1:6)); # x 是矩陣
x<-data.frame(x) # 不能是矩陣
f1<- function(col) {
   c(min(col),max(col))

}
vapply(x,f1,FUN.VALUE = c("min"=0,"max"=0)) # FUN.VALUE 是一個 named vector
```

```
    x1 x2
min  3  1
max  3  6
```

```
# vapply(x,f1,FUN.VALUE = c("min","max")) # 錯誤
```

2.20.6 mapply 函數

函數定義:

```
mapply(FUN, ..., MoreArgs = NULL, SIMPLIFY = TRUE,USE.NAMES = TRUE)
```

參數列表:

- FUN: 自定義的調用函數
- ...: 接收多個數據
- MoreArgs: 參數列表
- SIMPLIFY: 是否轉矩陣，當值 array 時，輸出結果按數組進行分組
- USE.NAMES: 如果 X 為字符串，TRUE 設置字符串為數據名，FALSE 不設置
- 如果 Fun=f(x)，則 mapply(f,a1,a2) 的執行是:
  f(a1), f(a2)
- 如果 Fun=f(x,y)，則 mapply(f,a1,a2) 的執行是:
  f(a1, a2)

比如，比較 3 個向量大小，按索引順序取較大的值。

```
set.seed(1)

# 定義 3 個向量
x<-1:10
y<-5:-4
z<-round(runif(10,-5,5))

# 按索引順序取較大的值。
mapply(max,x,y,z)
```

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

再看一個例子，生成 4 個符合正態分佈的數據集，分別對應的均值和方差為 c(1,10,100,1000)。

```
set.seed(1)

# 長度為 4
n<-rep(4,4)

# m 為均值，v 為方差
m<-v<-c(1,10,100,1000)

# 生成 4 組數據，按列分組
mapply(rnorm,n,m,v)
```

```
      [,1] [,2]  [,3]  [,4]
[1,] 0.374 13.3 157.6   379
[2,] 1.184  1.8  69.5 -1215
[3,] 0.164 14.9 251.2  2125
[4,] 2.595 17.4 139.0   955
```

2.20.7 tapply 函數

tapply 用於分組的循環計算，通過 INDEX 參數可以把數據集 X 進行分組，相當於 group by 的操作。

函數定義：

```
tapply(X, INDEX, FUN = NULL, ..., simplify = TRUE)
```

參數列表：

- X: 向量
- INDEX: 用於分組的索引
- FUN: 自定義的調用函數
- …: 接收多個數據
- simplify：是否數組化，當值 array 時，輸出結果按數組進行分組

比如，計算不同品種的鳶尾花的花瓣 (iris) 長度的均值。

```
# 通過 iris$Species 品種進行分組
tapply(iris$Petal.Length,iris$Species,mean)
```

```
    setosa versicolor  virginica
      1.46       4.26       5.55
```

分組求和的範例。

```
set.seed(1)
x<-1:10

# 亂衕分成 3 組，分別是 0,1,2
t<-round(runif(10,1,100)%%2);t
```

```
 [1] 1 2 2 1 1 2 1 0 1 1
```

```
# 對 x 進行分組求和
tapply(x,t,sum)
```

```
 0  1  2
 8 36 11
```

2.20.8 rapply 函數

2.20.9 eapply 函數

函數定義：

```
eapply(env, FUN, ..., all.names = FALSE, USE.NAMES = TRUE)
```

2.21 相關操作

practice1.R

2.21.1 問題: 想要知道有甚麼資料庫可以用?

例如，想要知道資料庫裏面有沒有’mtcars’
hint: grep(‘pattern’,target)

```
l<-list('a','b','c')
grep('a',l)
```

```
[1] 1
```

利用指令 `data()` 可以列出內建資料:

```
xx<-data()
```

xx 是甚麼? 直接打入 xx 並沒有像其他變數一樣，在 console 顯示內容:

```
xx
```

```
str(xx)
```

```
List of 4
 $ title  : chr "Data sets"
 $ header : NULL
 $ results: chr [1:138, 1:4] "forcats" "stringr" "stringr" "stringr" ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:4] "Package" "LibPath" "Item" "Title"
 $ footer : chr "Use 'data(package = .packages(all.available = TRUE))'\nto list the da
 - attr(*, "class")= chr "packageIQR"
```

```
is.data.frame(xx)
```

```
[1] FALSE
```

```
is.list(xx)
```

```
[1] TRUE
```

```
is.table(xx)
```

```
[1] FALSE
```

```
class(xx)
```

```
[1] "packageIQR"
```

`class()` 函數說明 xx 是一個 packageIQR 物件。`str()` 可以看出是一個長度為 4 的 list，其中 results 標籤，是一個字串矩陣，大小為 104 x 4。其中 4 個欄位的名稱為: “Package” “LibPath” “Item” “Title”（如下）:

```
head(xx$results)
```

```
     Package   LibPath
[1,] "forcats" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
[2,] "stringr" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
[3,] "stringr" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
[4,] "stringr" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
[5,] "ggplot2" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
[6,] "ggplot2" "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5"
     Item
[1,] "gss_cat"
[2,] "fruit"
[3,] "sentences"
[4,] "words"
[5,] "diamonds"
[6,] "economics"
     Title
[1,] "A sample of categorical variables from the General Social survey"
[2,] "Sample character vectors for practicing string manipulations."
[3,] "Sample character vectors for practicing string manipulations."
[4,] "Sample character vectors for practicing string manipulations."
[5,] "Prices of 50,000 round cut diamonds"
[6,] "US economic time series"
```

如果覺得混亂, 可以利用

```
View(xx$results)
```

因為是矩陣, 所以不能以「$」得到 Item 的內容?

```
xx$results$Item
```

```
Error in xx$results$Item:
  $ operator is invalid for atomic vectors
```

```
head(xx$results[,'item'])
```

```
Error in xx$results[, "item"]: 
  下標超出邊界
```

為甚麼錯誤?
因為大小寫, 所以正確寫法為:

```
head(xx$results[,'Item'])
```

```
[1] "gss_cat"   "fruit"     "sentences" "words"     "diamonds" 
[6] "economics"
```

利用 `grep()` 可以拿到 `mtcars` 的位置:

```
idx<-grep('mtcars',xx$results[,'Item'])
idx
```

```
[1] 106
```

確認

```
xx$results[idx,'Item']
```

```
    Item 
"mtcars" 
```

2.22 常見錯誤

2.22.1 error

```
l0 <- list(1, c(TRUE, FALSE), c("a", "b", "c"))
l0
```

> 錯誤念法
> l0 等於一個列表 list, 裡面有三個元素, 第一個元素是 1, 第二個元素是 c(TRUE, FALSE), 第三個元素是 c(“a”, “b”, “c”)

正確念法：
正確念法指派一個 list 給變數 l0，裡面有三個元素，第一個元素是純量 1，第二個元素是邏輯向量，長度為 2，內容分別是 TRUE, FALSE，第三個元素是字元向量，長度為 3，內容分別是 “a”, “b”, “c”

```r
names(l1) <- c("A","B","C")
```

錯誤念法
設定 names(li) 等於集合 (“A”,“B”,“C”)

正確念法：
將 l1 這個 list 的前 **3** 個元素名稱，分別指定為 A,B,C

2.22.2 error

```r
l1$x <- NULL
```

正確念法：
刪除元素名為 x 的元素。

2.22.3 error

```r
l0 <- list(1, c(TRUE, FALSE), c("a", "b", "c"))

l1 <- list(x = 1, y = c(TRUE, FALSE), z = c("a", "b", "c"))
l1$x
l1$y
l1$z
l1$m
```

更正念法建立 list 變數 `l1`, 名稱分別是 x,y,z, 內容分別是……
l1$yl1$z 為字元向量
尤其是這句
—-> ?? l1 y 產生兩次邏輯運算。?? <—- 邏輯運算指的是 and, or 因此，應該說 l1$y 是邏輯向量，元素有 2 個。

Chapter 3

Regular expression syntax

利用一些特殊字元,例如 $ * + . ? [ ] ^ { } | ( ) \構成 Regular expressions,用來在文字中表達某些搜尋樣式的語法。這裡只是簡短的介紹

3.0.0.1 Functions which work with regular expression patterns

- strsplit: split string into substrings at occurances of regexp
- grep: search for a regular expression within a string
- sub: search and then replace an occurance of a regular expression in a strng
- gsub: global search and replace all occurances of a regular expression in a string

3.0.1 Escape sequences

正規表達式中 . 代表任意字元, 但是如果我們要在字串中查找 “.”(例如, 檔案名稱中的點) 的時候要怎樣表示? 答案是使用逃逸字元\。但是在 R 語言中, 正規表達式，也是使用字串，可是在字串中,\本身就逃逸字元，因此需要\\.。

```r
dot <- "\\."
writeLines(dot)
```

```
\.
```

```r
x <- "a\\b"
writeLines(x)
```

```
a\b
```

3.0.1.1 grep, grepl

grep,grepl 的差別是後者不支援參數 value，且比對的結果和搜尋內容的個數相同。

```
grep("a\\.c", c("abc", "a.c", "bef"))
```

```
[1] 2
```

```
grep("a\\.c", c("abc", "a.c", "bef"),value=TRUE)
```

```
[1] "a.c"
```

```
grepl("a\\.c", c("abc", "a.c", "bef"))
```

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

又另一個例子, 是在字串中尋找單引號, 假定要在國家中找尋是否有名稱包含單引號的名字, 可以使用`'\''` 問題: 為甚麼不是`'\\''` hint: 這裡的逃逸字元只是字串的標準用法,`'`不是正規表達式的特殊字元。

```
country<-read.delim('resources/country.txt',header=FALSE)
names(country)<-"Name"
cname <-levels(country$Name)
grep('\'', cname, value = TRUE)
```

```
[1] "Cote d'Ivoire"
```

note: * 不是 country.names<-“Name”

問題: 下面兩行有甚麼差別?

```
grep('\'', country$Name, value = TRUE)
grep('\'', levels(country$Name), value = TRUE)
```

Hint: levels 傳回唯一值

```
levels(factor(c("a","b","a")))
```

```
[1] "a" "b"
```

```
factor(c("a","b","a"))
```

```
[1] a b a
Levels: a b
```

3.0.1.2 regexpr, gregexpr

regexpr 函數的使用方式類似 grepl。傳回整數向量 (表示符合位置，如果找不到，則傳回-1)，個數和比對字串串列相同。這個向量同時有屬性 match.length(參考範例)．

```
s<-"a.b.c.9"
regexpr('b',s )
```

```
[1] 3
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
regexpr('.',s ) # 要定位符號「.」, 這是錯的,\. 也錯
```

```
[1] 1
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
regexpr('\\.',s ) # 正確
```

```
[1] 2
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
regexpr('[.]',s ) # 正確 , 不需要 regexpr('[\\.]',s )
```

```
[1] 2
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
regexpr('\\d',s ) # 正確
```

```
[1] 7
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
x<-regexpr('[0-9]',s ) # 正確
attr(x,"match.length")
```

```
[1] 1
```

```
s<-"a.b.c.9 6 8"
x<-regexpr('[0-9]',s )
x
```

```
[1] 7
attr(,"match.length")
[1] 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
attr(x,"match.length")
```

```
[1] 1
```

gregexpr() 和 regexpr() 一樣, 不同的是會在目標字串中搜尋所有符合的位置而不是只找第一個 (g: 全局)。因此傳回的結果個數雖然和查找的串列元素個數一樣, 但是每個元素都是一個向量。如果都沒找到, 只有一個-1。

```
s<-"a.b.c.9 6 8"
 gregexpr('[0-9]',s )
```

```
[[1]]
[1]  7  9 11
attr(,"match.length")
[1] 1 1 1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

3.0.2 應用

產生 3 個檔案, 檔案名稱含有 cat

```
file.create(c("xcat1.rmd", "ycat2.txt", "zcat3.rmd"))
```

```
[1] TRUE TRUE TRUE
```

```
files<-list.files(".")
grep("cat", files, value = TRUE)
```

```
[1] "xcat1.rmd" "ycat2.txt" "zcat3.rmd"
```

```
grep("cat", files, value = FALSE)
```

```
[1] 64 66 67
```

```
grepl("cat", files)
```

```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[11] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[21] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[31] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[41] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[51] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[61] FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE
```

3.0.3 數量修飾詞 Quantifiers

指定重複次數.

- `*`: 至少 0 個。
- `+`: 至少一個.
- `?`: 最多一個.
- `{n}`: 剛好 n 個.
- `{n,}`: 至少 n 個.
- `{n,m}`: 次數在 n 到 m 之間 (包括 n,m).

```
(strings <- c("a", "ab", "acb", "accb", "acccb", "accccb"))
```

```
[1] "a"      "ab"     "acb"    "accb"   "acccb"  "accccb"
```

```
grep("ac*b", strings, value = TRUE)
```

```
[1] "ab"     "acb"    "accb"   "acccb"  "accccb"
```

```
grep("ac+b", strings, value = TRUE)
```

```
[1] "acb"    "accb"   "acccb"  "accccb"
```

```
grep("ac?b", strings, value = TRUE)
```

```
[1] "ab"  "acb"
```

```
grep("ac{2}b", strings, value = TRUE)
```

```
[1] "accb"
```

```
grep("ac{2,}b", strings, value = TRUE)
```

```
[1] "accb"   "acccb"  "accccb"
```

```r
grep("ac{2,3}b", strings, value = TRUE)
```

```
[1] "accb"  "acccb"
```

3.0.3.1 Exercise

Find all countries with `ee` in Gapminder using quantifiers.

```
[1] "Greece"
```

3.0.4 Position of pattern within the string

- `^`: 字串開頭.
- `$`: 字串結尾.
- `\b`: 左右都空白的 *word*.
- `\B`: 左右不是空白 *word*.

```r
(strings <- c("abcd", "cdab", "cabd", "c abd"))
```

```
[1] "abcd"  "cdab"  "cabd"  "c abd"
```

```r
grep("ab", strings, value = TRUE)
```

```
[1] "abcd"  "cdab"  "cabd"  "c abd"
```

```r
grep("^ab", strings, value = TRUE)
```

```
[1] "abcd"
```

```r
grep("ab$", strings, value = TRUE)
```

```
[1] "cdab"
```

```r
grep("\\bab", strings, value = TRUE)
```

```
[1] "abcd"  "c abd"
```

```
regexpr("\\bab", strings)
```

```
[1]  1 -1 -1  3
attr(,"match.length")
[1]  2 -1 -1  2
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

```
regexpr("\\Bab", strings)
```

```
[1] -1  3  2 -1
attr(,"match.length")
[1] -1  2  2 -1
attr(,"index.type")
[1] "chars"
attr(,"useBytes")
[1] TRUE
```

3.0.4.1 Exercise

Find all `.txt` files in the repository.

```
[1] "ycat2.txt"
```

```
[1] "ycat2.txt"
```

3.0.5 Operators

- `.`：任意字元。
- `[...]`：例如,[ade] 3 個字元其中一個。[a-e],a 到 e.
- `[^...]`：除了指定的字元，非 [...] 的意思。.
- `\`：抑制下列字元在字串,`""`, 中的特殊意思 `$ * + . ? [ ] ^ { } | ( ) \`, 只是在 R 中, 因為\又有特殊意義, 因此必須雙\例如 `\\$`。
- `|`：或.
- `(...)`：字元組, 根據出現的順序, 可以用 `\\N` 識別 (**backreference**).

```
(strings <- c("^ab", "ab", "abc", "abd", "abe", "ab 12"))
```

```
[1] "^ab"   "ab"    "abc"   "abd"   "abe"   "ab 12"
```

```
grep("ab.", strings, value = TRUE)
```

```
[1] "abc"   "abd"   "abe"   "ab 12"
```

```
grep("ab[c-e]", strings, value = TRUE)
```

```
[1] "abc" "abd" "abe"
```

```
grep("ab[^c]", strings, value = TRUE)
```

```
[1] "abd"   "abe"   "ab 12"
```

```
grep("^ab", strings, value = TRUE)
```

```
[1] "ab"    "abc"   "abd"   "abe"   "ab 12"
```

```
grep("\\^ab", strings, value = TRUE)
```

```
[1] "^ab"
```

```
grep("abc|abd", strings, value = TRUE)
```

```
[1] "abc" "abd"
```

```
gsub("(ab) 12", "\\1 34", strings)
```

```
[1] "^ab"   "ab"    "abc"   "abd"   "abe"   "ab 34"
```

note: 要讓. 失效，可以 [.] 或 \\. 表示。

3.0.5.1 Excercise

找出國名中間有字元 i t 並且將 land 結尾的字串改為大寫。例如,Finland->FinLAND

```
[1] "FinLAND"     "IceLAND"     "IreLAND"     "SwaziLAND"  
[5] "SwitzerLAND" "ThaiLAND"
```

3.0.6 Character classes

預設字元分類例如 numbers, letters, 分類。開頭 `[:` 結尾 `:]` ，另一種簡寫方式為利用 `\` 。

- `[:digit:]` or `\d`: 數字, 等同 `[0-9]`.
- `\D`: 非數字, 等同 `[^0-9]`.
- `[:lower:]`: 小寫, 等同 `[a-z]`.
- `[:upper:]`: 大寫, 等同 `[A-Z]`.
- `[:alpha:]`: 字母, 等同 `[[:lower:][:upper:]]` or `[A-z]`.
- `[:alnum:]`: 等同 `[[:alpha:][:digit:]]` 或 `[A-z0-9]`.
- `\w`: word characters, equivalent to `[[:alnum:]_]` or `[A-z0-9_]`.
- `\W`: not word, equivalent to `[^A-z0-9_]`.
- `[:xdigit:]`: hexadecimal digits (base 16), 0 1 2 3 4 5 6 7 8 9 A B C D E F a b c d e f, equivalent to `[0-9A-Fa-f]`.
- `[:blank:]`: blank characters, i.e. space and tab.
- `[:space:]`: space characters: tab, newline, vertical tab, form feed, carriage return, space.
- `\s`: space, .
- `\S`: not space.
- `[:punct:]`: punctuation characters, ! ” # $ % & ’ ( ) * + , - . / : ; < = > ? @ [ ] ˆ _ ‘ { | } ~.
- `[:graph:]`: graphical (human readable) characters: equivalent to `[[:alnum:][:punct:]]`.
- `[:print:]`: printable characters, equivalent to `[[:alnum:][:punct:]\\s]`.
- `[:cntrl:]`: control characters, like `\n` or `\r`, `[\x00-\x1F\x7F]`.

Note:

- `[:...:]` 必須寫在中括號中, 例如 `[[:digit:]]`.
- `\` 注意 `\\d`. 和 `\t` 的區別。

3.1 attribute

oth the names and the dimensions of matrices and arrays are stored in R as attributes of the object. These attributes can be seen as labeled values you can attach to any object.

They form one of the mechanisms R uses to define specific object types like dates, time series, and so on. They can include any kind of information, and you can use them yourself to add information to any object.

To see all the attributes of an object, you can use the attributes() function. You can see all the attributes of my.array like this:

> attributes(my.array) $dim [1] 3 4 2 This function returns a named list, where each item in the list is an attribute. Each attribute can, on itself, be a list again. For example, the attribute dimnames is actually a list containing the row and column names of a matrix.

You can check that for yourself by checking the output of attributes(baskets.team). You can set all attributes as a named list as well. You find examples of that in the Help file ?attributes.

To get or set a single attribute, you can use the attr() function. This function takes two important arguments. The first argument is the object you want to examine, and the second argument is the name of the attribute you want to see or change. If the attribute you ask for doesn't exist, R simply returns NULL.

Imagine you want to add which season Granny and Geraldine scored the baskets mentioned in baskets.team. You can do this with the following code:

> attr(baskets.team,'season') <- '2010-2011' To get the value of this attribute returned, you can then use following code:
>
> attr(baskets.team,'season') [1] "2010-2011" You can delete attributes again by setting their value to NULL, like this:
>
> attr(baskets.team,'season') <- NULL

Chapter 4

File System

暫時子目錄

函數 tempfile() 不是建立新檔案, 而是在目前的 r session 中隨機產生唯一檔案名稱，檔案位置預設是在暫時子目錄中。

```
mydirname <- tempfile(pattern = "mydir")
```

mydirname [1] ”C:\Users\linchao\AppData\Local\Temp\RtmpIp3ZiD

4.1 Exploring file system

function file.exists() 可以用來知道檔案是否存在, function dir() 用來知道目前檔案位置的內容.

```
file.exists(mydirname)
```

dir(tempdir()) character(0) # Empty character vector ::: sidebar a <- character(0) identical(a, character(0)) # returns TRUE

identical(a, ””) # returns FALSE identical(a, numeric(0)) # returns also FALSE hint: 利用 length :::

4.2 Creating of a directory

dir.create 建立子目錄

```
dir.create(mydirname)
file.exists(mydirname) # 上面指令建立的子目錄是否存在
dir(tempdir(), full.names = TRUE) # 列出目前子目錄內容 (全名)
file.mtime(mydirname) # 子目錄建立時間,make time
```

```
[1] TRUE
[1] "C:\Users\linchao\AppData\Local\Temp\RtmpIp3ZiD/file87e8755a1876"
[2] "C:\Users\linchao\AppData\Local\Temp\RtmpIp3ZiD/mydir87e86b51384d"
[3] "C:\Users\linchao\AppData\Local\Temp\RtmpIp3ZiD/rs-graphics-0f3f81af-32b7-49c4-a272-ad1a859f222f"
[1] "2018-10-25 23:42:51 CST"
```

4.3 R 系統檔案列表

如果要觀察安裝套件的檔案內無, 可以使用指令 `system.file()`, 這個指令可以列出套建的全路徑。例如,

```
system.file(package = "stats")
```

```
[1] "C:/PROGRA~1/R/R-3.5.1/library/stats"
```

列出套件 stats 中, 所有的檔案

```
dir(system.file(package = "stats"))
```

```
[1] "COPYRIGHTS.modreg" "demo" "DESCRIPTION"
[4] "help" "html" "INDEX"
[7] "libs" "Meta" "NAMESPACE"
[10] "R" "SOURCES.ts"
```

上面可以看到套件中包含子目錄 `demo`, 如果還要查看 `demo` 中的檔案內容:

```
dir(system.file("demo", package = "stats"))
```

```
[1] "glm.vr.R" "lm.glm.R" "nlm.R" "smooth.R"
```

如果要看全路徑, 可以:

```
dir(system.file("demo", package = "stats"), full.names = TRUE)
```

```
[1] "/usr/lib64/R/library/stats/demo/glm.vr.R" [2] "/usr/lib64/R/library/stats/demo/lm.glm.R"
[3] "/usr/lib64/R/library/stats/demo/nlm.R"
[4] "/usr/lib64/R/library/stats/demo/smooth.R"
```

4.4 在 R 中建立檔案名稱

甚麼是檔案名稱? 不是 **character string** 而是 **system-specific character string**。例如,R 函數 `file.path()` 的執行結果就是 `system-specific file names`.

```
workingdir <- "projects"
projectdir <- "warandpeace"
datadir    <- "data"
file.path(workingdir, projectdir, datadir)
```

[1] “projects/warandpeace/data” 上面是 windows 10 系統上的結果, 比較 Linux-based OS, 則為:
[1] “projects/warandpeace/data” 注意斜線都一樣。

4.4.1 Note about Windows

利用函數 `file.path()` 列出的檔案路徑分隔是斜線, 而不是反斜線 (例如 for example `C:\Program Files\R\R-3.5.1`)?

求助文件說:

The components are by default separated by ‘/’ (not ‘’) on Windows. Note: DOS 和 Windows 也支援斜線.

4.4.2 和工作區相關的指令

dir(), list.files, list.dirs

getwd() setwd()

```
list.files(R.home())
## Only files starting with a-l or r
## Note that a-l is locale-dependent, but using case-insensitive
## matching makes it unambiguous in English locales
dir("../..", pattern = "^[a-lr]", full.names = TRUE, ignore.case = TRUE)

list.dirs(R.home("doc"))
list.dirs(R.home("doc"), full.names = FALSE)
```

### 應用範例

4.4.2.1 列出目前工作目錄上的檔案

```
x <- dir()
for (item in x) {
    show(item) #print(item)
}
```

或者, 僅列出前幾個檔案

```
x<-dir()
head(x)
```

4.5 讀檔

4.5.0.1 僅僅查看檔案內容

測試中文的檔案
測試檔案

注意檔案內容第一行不是資料。

`file.show("resources/wh.csv")` 會打開 excel
也可以利用殼層指令:
`system("cat resources/wh.csv")`, 但是如果內容是中文, 顯示結果會變成亂碼。

```
cat(readChar("resources/wh.csv", 1e5))
```

```
# this is a test file
height,weight,sex
156,56, m
167,., f
189,70, m
180,, f
```

利用 useBytes 不會出現警告

```
cat(readChar("resources/wh.csv", useBytes=TRUE,1e5))
```

```
# this is a test file
height,weight,sex
156,56, m
167,., f
189,70, m
180,, f
```

參數有 useBytes, 但是不管 TRUE/FALSE，都是中文亂碼

```
res <- readLines(system.file("DESCRIPTION", package="MASS"))
length(res)
```

4.5.1 一般的開檔/關檔過程

大體上每個語言的開檔和關檔案的過程，都會利用到作業系統，因此會有開檔動作，從作業系統得到一個數字，這裡稱為檔案指標，然後利用這個指標，讀入文字或寫入文字，然後有一個關閉檔案的動作告訴作業系統，處理檔案結束。

在 R 語言中，開檔不是 open(), 而是 file:

```
  con = file(filepath, "r")
  開始讀寫操作
  close(con)
```

4.5.1.1 csv

```
read.csv("filename.csv", #name of file
         header = TRUE, #are there column names in 1st row?
         sep = ",", #what separates rows?
         as.is = !stringsAsFactors, # 關掉字元轉 factor
         colClasses = NA # to convert everything to character data set to "character"
         na.string = "NA" # could be "." for SAS files
         skip = 0, # 要跳過的前幾行數目>0
         strip.white = TRUE, # 擠掉空白, 例如 " 0.1" = "0.1"
         fill = TRUE, #fill in rows that have unequal numbers of columns
         comment.char = "#", # 註解不讀入
         stringsAsFactors = FALSE # 比 as.is 常用, 關掉字元轉 factor
         )
```

read.csv() 最簡單的用法是直接給檔案名稱, 例如

```
rst = read.csv('resources/wh.csv')
```

```
Error in read.table(file = file, header = header, sep = sep, quote = quote, :
  more columns than column names

Error in      dec = dec, fill = fill, comment.char = comment.char, ...):
  more columns than column names
```

但是 wh.csv 前面有幾行是註解

```
readLines('resources/wh.csv',n=3)
```

```
[1] "# this is a test file" "height,weight,sex"
[3] "156,56, m"
```

因此加上 skip 參數

```
rst = read.csv('resources/wh.csv',skip=1)
head(rst)
```

```
  height weight sex
1    156     56   m
2    167      .   f
3    189     70   m
4    180          f
```

也可以利用參數 `comment.char`

```
rst = read.csv('resources/wh.csv',comment.char="#")
head(rst)
```

```
  height weight sex
1    156     56   m
2    167      .   f
3    189     70   m
4    180          f
```

但是注意到 `weight` 這個欄位被轉為字串, 因為有一個 “.” 。有些套裝軟體的 CSV 輸出將 NA 轉為 “.”, 因此

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
head(rst)
```

```
  height weight sex
1    156     56   m
2    167     NA   f
3    189     70   m
4    180     NA   f
```

4.6 write csv

4.6.1 data.frame to CSV

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv")
readLines("MyData.csv",n=3) # note: readLines() 不是 readline()
```

```
[1] "\"\",\"height\",\"weight\",\"sex\""
[2] "\"1\",156,56,\" m\""
[3] "\"2\",167,NA,\" f\""
```

如果 row names 不要寫進 csv, 可以利用參數 `row.names=FALSE`。

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE)
readLines("MyData.csv",n=3)
```

```
[1] "\"height\",\"weight\",\"sex\"" "156,56,\" m\""
[3] "167,NA,\" f\""
```

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE)
readLines("MyData.csv",n=3)
```

輸出是: [1] “”height”,”weight”,”sex””
[2] “156,56,” m””
[3] “167,NA,” f””

可以看出來每一行都有雙引號表示字串。
對照文字檔 (mData.csv)

“height”,“weight”,“sex”
156,56,” m”
167,NA,” f”
189,70,” m”
180,NA,” f”

1. 上面 sidebar 中, 顯示的的 sex 欄位, 可以看見” m” 也就是有空白, 問利用 `read.csv` 讀入時, 會是甚麼情況?

如果要指定 NA 的值, 可以利用欄位 =””, 否則預定是 `NA`

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE, na="")
cat(readLines("MyData.csv"), sep = '\n')
```

```
"height","weight","sex"
156,56," m"
167,," f"
189,70," m"
180,," f"
```

Note: 參考 rmarkdown 提要::: todo cat :::

如果輸出時, 字串不要雙引號, 則可以用 `quote=FALSE`

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE, quote=F)
# system("cat myData.csv") # 這個不行
cat(readLines('MyData.csv'), sep = '\n')
```

```
height,weight,sex
156,56, m
167,NA, f
189,70, m
180,NA, f
```

如果要更多的輸出控制可以參考可以`write.table()`. `write.csv()` 函數, 實際上會再呼叫 `write.table()` 函數。例如, 欄位名稱也不要輸出的話, 可以利用 `write.table()`, 要設定的參數如下: `sep=”,”` 和 `col.names=FALSE`, 前者是 CSV 的主要分隔字元, 後者則是不要欄位名稱。

```
write.table(rst, file = "MyData.csv",row.names=FALSE, na="",col.names=FALSE, sep=",")
cat(readLines('MyData.csv'), sep = '\n')
```

```
156,56," m"
167,," f"
189,70," m"
180,," f"
```

4.7 R 語言的檔案格式

如果不考慮存檔格式, 那麼可以利用

```
save(v1,v2,...file="myfile.RData")
...
load(file="myfie.RData")
```

Chapter 5

Functions

5.1 Introduction

簡單測試 ### Quiz {-}

Answer the following questions to see if you can safely skip this chapter. You can find the answers in ??.

1. What are the three components of a function?

2. What does the following code return?

   ```
   x <- 10
   f1 <- function(x) {
     function() {
       x + 10
     }
   }
   f1(1)()
   ```

3. How would you usually write this code?

   ```
   `+`(1, `*`(2, 3))
   ```

4. How could you make this call easier to read?

   ```
   mean(, TRUE, x = c(1:10, NA))
   ```

5. Does the following code throw an error when executed? Why/why not?

```
f2 <- function(a, b) {
  a * 10
}
f2(10, stop("This is an error!"))
```

6. What is an infix function? How do you write it? What's a replacement function? How do you write it?

7. How do you ensure that cleanup action occurs regardless of how a function exits?

Outline

- Section 5.2 describes the basics of creating a function, the three main components of a function, and the exception to many function rules: primitive functions (which are implemented in C, not R).

- Section 5.3 shows you how R finds the value associated with a given name, i.e. the rules of lexical scoping.

- Section 5.4 is devoted to an important property of function arguments: they are only evaluated when used for the first time.

- Section 5.6 discusses the two primary ways that a function can exit, and how to define an exit handler, code that is run on exit, regardless of what triggers it.

- Section 5.7 shows you the various ways in which R disguises ordinary function calls, and how you can use the standard prefix form to better understand what's going on.

5.2 Function fundamentals

幾個重要觀念

- 函數也是物件, 就像是 vectors 也是物件。

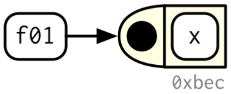
- 由三個部份組成: arguments, body, and environment.

There are exceptions to every rule, and in this case, there is a small selection of "primitive" base functions that are implemented purely in C.

5.2.1 First-class functions

在 R 中, 函數也是物件, 這種特性也叫做 “first-class functions”. 如下:

```r
f01 <- function(x) {
  sin(1 / x ^ 2)
}
```



匿名函數:

```r
lapply(mtcars, function(x) length(unique(x)))
Filter(function(x) !is.numeric(x), mtcars)
integrate(function(x) sin(x) ^ 2, 0, pi)
```

在 list 中, 也可以放入:

```r
funs <- list(
  half = function(x) x / 2,
  double = function(x) x * 2
)

funs$double(10)
```

```
[1] 20
```

在 R 語言中, 函數有叫做 closure 因為,R 函數包含 (enclose) 它們的環境 environments.

```r
typeof(f01)
```

```
[1] "closure"
```

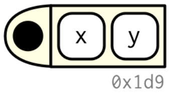
5.2.2 Function components

1 個函數有 3 個部分:

- `formals()`, 參數

- `body()`, {} 內部.
- `environment()`, 決定函數怎樣找出變數 (names) 的內容。.

I'll draw functions as in the following diagram. The black dot on the left is the environment. The two blocks to the right are the function arguments. I won't draw the body, because it's usually large, and doesn't help you understand the "shape" of the function.



The function environment always exists, but it is only printed when the function isn't defined in the global environment.

```r
f02 <- function(x) {
  # A comment
  x ^ 2
}

formals(f02)
```

```
$x
```

```r
body(f02)
```

```
{
    x^2
}
```

```r
environment(f02)
```

```
<environment: R_GlobalEnv>
```

就像所以其他 R 的物件, 函數也有很多 `attributes()`. 其中一個 "srcref", 是 source reference 的縮寫。

```r
attr(f02, "srcref")
```

```
function(x) {
  # A comment
  x ^ 2
}
```

5.2.3 Primitive functions

3 個組件的規則有例外，像是 Primitive functions, like `sum()` and `[`, 直接調用 C 語言。

```
sum
```

```
function (..., na.rm = FALSE)  .Primitive("sum")
```

```
`[`
```

```
.Primitive("[")
```

看一下 type 分別屬於 “builtin” or “special”:

```
typeof(sum)
```

```
[1] "builtin"
```

```
typeof(`[`)
```

```
[1] "special"
```

因此，`formals()`, `body()`, and `environment()` 都回傳 `NULL`:

```
formals(sum)
```

```
NULL
```

```
body(sum)
```

```
NULL
```

```
environment(sum)
```

```
NULL
```

這些所謂的原始函數，只存在於基本套件 (base packages) 。.

5.2.4 Exercises

1. Given a function, like `"mean"`, `match.fun()` lets you find a function. Given a function, can you find its name? Why doesn't that make sense in R?

2. It's possible (although typically not useful) to call an anonymous function. Which of the two approaches below is correct? Why?

   ```
   function(x) 3()
   ```

   ```
   function(x) 3()
   ```

   ```
   (function(x) 3)()
   ```

   ```
   [1] 3
   ```

3. A good rule of thumb is that an anonymous function should fit on one line and shouldn't need to use `{}`. Review your code. Where could you have used an anonymous function instead of a named function? Where should you have used a named function instead of an anonymous function?

4. What function allows you to tell if an object is a function? What function allows you to tell if a function is a primitive function?

5. This code makes a list of all functions in the base package.

   ```
   objs <- mget(ls("package:base"), inherits = TRUE)
   funs <- Filter(is.function, objs)
   ```

   Use it to answer the following questions:

   a. Which base function has the most arguments?
   b. How many base functions have no arguments? What's special about those functions?
   c. How could you adapt the code to find all primitive functions?

6. What are the three important components of a function?

7. When does printing a function not show the environment it was created in?

5.3 Lexical scoping

In [Names and values], we discussed assignment, the act of binding a name to a value. Here we'll discuss **scoping**, the act of finding the value associated with a name.

下面的執行結果傳回 10 還是 20?[1]

```r
x <- 10
g01 <- function() {
  x <- 20
  x
}

g01()
```

了解範圍規則, 有助於函數的模組開發, 甚至有助於將 R 翻譯到其他語言。

lexical scoping [2]: it looks up the values of names based on how a function is defined, not how it is called. "Lexical" here is not the English adjective "relating to words or a vocabulary". It's a technical CS term that tells us that the scoping rules use a parse-time, rather than a run-time structure.

R 的's lexical scoping 遵循 4 個主要規則::

- Name masking
- Functions vs. variables
- A fresh start
- Dynamic lookup

5.3.1 Name masking

內部範圍的宣告 (第一次使用) 覆蓋外部範圍的宣告。.

```r
x <- 10
y <- 20
g02 <- function() {
  x <- 1
  y <- 2
  c(x, y)
}
g02()
```

[1] 20.

[2] Functions that automatically quote one or more arguments (sometimes called NSE functions) can override the default scoping rules to implement other varieties of scoping. You'll learn more about that in metaprogramming.

```
[1] 1 2
```

如果在內部宣告找不到, 就找外一層，一直到 global environment。

```
x <- 2
g03 <- function() {
  y <- 1
  c(x, y)
}
g03()
```

```
[1] 2 1
```

上面的規則仍然適用於函數中的函數.

測試: 下面的 R 程式會有甚麼結果? [3]

```
x <- 1
g04 <- function() {
  y <- 2
  i <- function() {
    z <- 3
    c(x, y, z)
  }
  i()
}
g04()
```

同樣也適用於建立函數的函數 (**closures**). 參考 [closures]; 這裡只是用來說明上述規則的使用。`g05()`, 傳回函數, 猜猜執行結果?[4]

```
x <- 10
y <- 20

g05 <- function() {
  y <- 2
  function() {
    c(x, y)
  }
}
g06 <- g05()
g06()
```

[3] `g04()` returns `c(1, 2, 3)`.

[4] `g06()` returns `c(10, 2)`.

This seems a little magical: how does R know what the value of y is after j() has returned? It works because k preserves the environment in which it was defined and because the environment includes the value of y. You'll learn more about how environments work in Environments.

5.3.2 Functions vs. variables

既然函數也只是普通的物件, 那麼同樣的名稱尋找規則也適用於函數: 這個例子中,g07 在外部和內部皆有定義。

```
g07 <- function(x) x + 1
g08 <- function() {
  g07 <- function(x) x + 100
  g07(10)
}
g08()
```

```
[1] 110
```

但是如果同一個名稱, 在不同範圍有不一樣的型態呢? 例如 g9 一個是變數, 一個是函數:

```
g09 <- function(x) x + 100
g10 <- function() {
  g09 <- 10
  g09(g09)
}
g10()
```

```
[1] 110
```

一般來講, 上面的用法在語法上是沒問題, 但是最好避免。

5.3.3 A fresh start

第一次執行和第二次執行有甚麼不同?[5]

函數 exists(x)：會尋找變數名稱 x 是否存在, 存在則無回 TRUE, 否則傳回 FALSE.)

[5] g11() 每次被調用都是傳回 1。.

```
g11 <- function() {
  if (!exists("a")) {
    a <- 1
  } else {
    a <- a + 1
  }
  a
}

g11()
g11()
```

每次執行的時候, 一個新的 environment 會被建立, 用來主導函數的執行。**??**.)

5.3.4 Dynamic lookup

Lexical scoping determines where to look for values, not when to look for them. R looks for values when the function is run, not when it's created. This means that the output of a function can differ depending on objects outside its environment:

```
g12 <- function() x + 1
x <- 15
g12()
```

```
[1] 16
```

```
x <- 20
g12()
```

```
[1] 21
```

This behaviour can be quite annoying. If you make a spelling mistake in your code, you won't get an error when you create the function, and you might not even get one when you run the function, depending on what variables are defined in the global environment.

One way to detect this problem is to use `codetools::findGlobals()`. This function lists all the external dependencies (unbound symbols) within a function:

```
codetools::findGlobals(g12)
```

```
[1] "+" "x"
```

Another way to solve the problem would be to manually change the environment of the function to the `emptyenv()`, an environment which contains nothing:

```
environment(g12) <- emptyenv()
g12()
```

```
Error in x + 1:
  沒有這個函數 "+"
```

Both of these approaches reveal why this undesirable behaviour exists: R relies on lexical scoping to find *everything*, even the `+` operator. This provides a rather beautiful simplicity to R's scoping rules.

5.3.5 Exercises

1. What does the following code return? Why? Describe how each of the three `c`'s is interpreted.

   ```
   c <- 10
   c(c = c)
   ```

2. What are the four principles that govern how R looks for values?

3. What does the following function return? Make a prediction before running the code yourself.

   ```
   f <- function(x) {
     f <- function(x) {
       f <- function() {
         x ^ 2
       }
       f() + 1
     }
     f(x) * 2
   }
   f(10)
   ```

5.4 Lazy evaluation

In R, function arguments are **lazily evaluated**: they're only evaluated if accessed. For example, this code doesn't generate an error because `x` is never used:

```
h01 <- function(x) {
  10
}
h01(stop("This is an error!"))
```

```
[1] 10
```

This is an important feature because it allows you to do things like include potentially expensive computations in function arguments that will only be evaluated if needed.

5.4.1 Forcing evaluation

To **compel** the evaluation of an argument, use `force()`:

```
h02 <- function(x) {
  force(x)
  10
}
h02(stop("This is an error!"))
```

```
Error in force(x):
  This is an error!
```

It is usually not necessary to force evaluation. It's needed primarily for certain functional programming techniques which we'll cover in detail in [function operators]. Here, I want to show you the basic issue.

Take this small but surprisingly tricky function. It takes a single argument `x`, and returns a function that returns `x` when called.

```
capture1 <- function(x) {
  function() {
    x
  }
}
```

There's a subtle issue with this function: the value of `x` will be captured not when you call `capture()`, but when you call the function that `capture()` returns:

```
x <- 10
h03 <- capture1(x)
h04 <- capture1(x)

h03()
```

```
[1] 10
```

```r
x <- 20
h04()
```

```
[1] 20
```

Even more confusingly this only happens once: the value is locked in after you have called `h03()`/`h04()` for the first time.

```r
x <- 30
h03()
```

```
[1] 10
```

```r
h04()
```

```
[1] 20
```

This behaviour is a consequence of lazy evaluation. The `x` argument is evaluated once `h03()`/`h04()` is called, and then its value is cached. We can avoid the confusion by forcing `x`:

```r
capture2 <- function(x) {
  force(x)

  function() {
    x
  }
}

x <- 10
h05 <- capture2(x)

x <- 20
h05()
```

```
[1] 10
```

5.4.2 Promises

}

Lazy evaluation is powered by a data structure called a **promise**, or (less commonly) a thunk. We'll come back to this data structure in metaprogramming because it's one of the features of R that makes it most interesting as a programming language.

A promise has three components:

- The expression, like `x + y` which gives rise to the delayed computation.
- The environment where the expression should be evaluated.
- The value, which is computed and cached when the promise is first accessed by evaluating the expression in the specified environment.

The value cache ensures that accessing the promise multiple times always returns the same value. For example, you can see in the following code that `runif(1)` is only evaluated once:

```
h06 <- function(x) {
  c(x, x, x)
}

h06(runif(1))
```

```
[1] 0.0808 0.0808 0.0808
```

You can also create promises "by hand" using `delayedAssign()`:

```
delayedAssign("x", {print("Executing code"); runif(1)})
x
```

```
[1] "Executing code"
```

```
[1] 0.834
```

```
x
```

```
[1] 0.834
```

You'll see this idea again in advanced bindings.

5.4.3 Default arguments

Thanks to lazy evaluation, default value can be defined in terms of other arguments, or even in terms of variables defined later in the function:

```r
h07 <- function(x = 1, y = x * 2, z = a + b) {
  a <- 10
  b <- 100

  c(x, y, z)
}

h07()
```

```
[1]   1   2 110
```

Many base R functions use this technique, but I don't recommend it. It makes code harder to understand because it requires that you know exactly *when* default arguments are evaluated in order to predict *what* they will evaluate to.

The evaluation environment is slightly different for default and user supplied arguments, as default arguments are evaluated inside the function. This means that seemingly identical calls can yield different results. It's easiest to see this with an extreme example:

```r
h08 <- function(x = ls()) {
  a <- 1
  x
}

# ls() evaluated inside f:
h08()
#> [1] "a" "x"

# ls() evaluated in global environment:
h08(ls())
#> [1] "f"
```

5.4.4 Missing arguments

If an argument has a default, you can determine if the value comes from the user or the default with `missing()`:

```
h09 <- function(x = 10) {
  list(missing(x), x)
}
str(h09())
```

```
List of 2
 $ : logi TRUE
 $ : num 10
```

```
str(h09(10))
```

```
List of 2
 $ : logi FALSE
 $ : num 10
```

`missing()` is best used sparingly. Take `sample()`, for example. How many arguments are required?

```
args(sample)
```

```
function (x, size, replace = FALSE, prob = NULL)
NULL
```

It looks like both `x` and `size` are required, but in fact `sample()` uses `missing()` to provide a default for `size` if it's not supplied. If I was to rewrite sample myself[6], I'd use an explicit `NULL` to indicate that `size` can be supplied, but it's not required:

```
sample <- function(x, size = NULL, replace = FALSE, prob = NULL) {
  if (is.null(size)) {
    size <- length(x)
  }

  x[sample.int(length(x), size, replace = replace, prob = prob)]
}
```

You can make that pattern even simpler with a small helper. The infix `%||%` function uses the LHS if it's not null, otherwise it uses the RHS:

[6] Note that this only implements one way of calling `sample()`: you can also call it with a single integer, like `sample(10)`. This unfortunately makes `sample()` prone to silent errors in situations like `sample(x[i])`.

```
`%||%` <- function(lhs, rhs) {
  if (!is.null(lhs)) {
    lhs
  } else {
    rhs
  }
}

sample <- function(x, size = NULL, replace = FALSE, prob = NULL) {
  size <- size %||% length(x)
  x[sample.int(length(x), size, replace = replace, prob = prob)]
}
```

Because of lazy evaluation, you don't need to worry about unnecessary computation: the RHS of `%||%` will only be evaluated if the LHS is null.

5.4.5 Exercises

1. What important property of `&&` make `x_ok()` work?

   ```
   x_ok <- function(x) {
     !is.null(x) && length(x) == 1 && x > 0
   }

   x_ok(NULL)
   ```

   ```
   [1] FALSE
   ```

   ```
   x_ok(1)
   ```

   ```
   [1] TRUE
   ```

   ```
   x_ok(1:3)
   ```

   ```
   [1] FALSE
   ```

   What is different with this code? Why is this behaviour undesirable here?

   ```
   x_ok <- function(x) {
     !is.null(x) & length(x) == 1 & x > 0
   }

   x_ok(NULL)
   ```

```
logical(0)
```

```r
x_ok(1)
```

```
[1] TRUE
```

```r
x_ok(1:3)
```

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

2. The definition of `force()` is simple:

   ```r
   force
   ```

   ```
   function (x)
   x
   <bytecode: 0x55a96ac06ce8>
   <environment: namespace:base>
   ```

   Why is it better to `force(x)` instead of just `x`?

3. What does this function return? Why? Which principle does it illustrate?

   ```r
   f2 <- function(x = z) {
     z <- 100
     x
   }
   f2()
   ```

4. What does this function return? Why? Which principle does it illustrate?

   ```r
   y <- 10
   f1 <- function(x = {y <- 1; 2}, y = 0) {
     c(x, y)
   }
   f1()
   y
   ```

5. In `hist()`, the default value of `xlim` is `range(breaks)`, the default value for `breaks` is `"Sturges"`, and

   ```r
   range("Sturges")
   ```

   ```
   [1] "Sturges" "Sturges"
   ```

Explain how `hist()` works to get a correct `xlim` value.

6. Explain why this function works. Why is it confusing?

```
show_time <- function(x = stop("Error!")) {
  stop <- function(...) Sys.time()
  print(x)
}
show_time()
```

```
[1] "2018-11-19 21:50:48 CST"
```

7. How many arguments are required when calling `library()`?

5.5 ... (dot-dot-dot)

Functions can have a special argument `...` (pronounced dot-dot-dot). If a function has this argument, it can take any number of additional arguments. In other programming languages, this type of argument is often called a varargs, or the function is said to be variadic.

Inside a function, you can use `...` to pass those additional arguments on to another function:

```
i01 <- function(y, z) {
  list(y = y, z = z)
}

i02 <- function(x, ...) {
  i01(...)
}

str(i02(x = 1, y = 2, z = 3))
```

```
List of 2
 $ y: num 2
 $ z: num 3
```

It's possible (but rarely useful) to refer to elements of `...` by their position, using a special form:

```
i03 <- function(...) {
  list(first = ..1, third = ..3)
}
str(i03(1, 2, 3))
```

```
List of 2
 $ first: num 1
 $ third: num 3
```

More often useful is `list(...)`, which evaluates the arguments and stores them in a list:

```r
i04 <- function(...) {
  list(...)
}
str(i04(a = 1, b = 2))
```

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

(See also `rlang::list2()` to support splicing and to silently ignore trailing commas, and `rlang::enquos()` to capture the unevaluated arguments, the topic of [quasiquotation].)

There are two primary uses of `...`, both of which we'll come back to later in the book:

- If your function takes a function as an argument, you want some way to pass on additional arguments to that function. In this example, `lapply()` uses `...` to pass `na.rm` on to `mean()`:

  ```r
  x <- list(c(1, 3, NA), c(4, NA, 6))
  str(lapply(x, mean, na.rm = TRUE))
  ```

  ```
  List of 2
   $ : num 2
   $ : num 5
  ```

  We'll come back to this technique in Section **??**.

- If your function is an S3 generic, you need some way to allow methods to take arbitrary extra arguments. For example, take the `print()` function. There are different options for printing types of object, so there's no way for the print generic to prespecify every possible argument. Instead, it uses `...` to allow individual methods to have different arguments:

  ```r
  print(factor(letters), max.levels = 4)

  print(y ~ x, showEnv = TRUE)
  ```

We'll come back to this use of `...` in Section **??**.

Using `...` comes with two downsides:

- When you use it to pass arguments on to another function, you have to carefully explain to the user where those arguments go. This makes it hard to understand the what you can do with functions like `lapply()` and `plot()`.
- Any misspelled arguments will not raise an error. This makes it easy for typos to go unnoticed:

```
sum(1, 2, NA, na_rm = TRUE)
```

```
[1] NA
```

`...` is a powerful tool, but be aware of the downsides.

5.5.1 Exercises

1. Explain the following results:

```
sum(1, 2, 3)
```

```
[1] 6
```

```
mean(1, 2, 3)
```

```
[1] 1
```

```
sum(1, 2, 3, na.omit = TRUE)
```

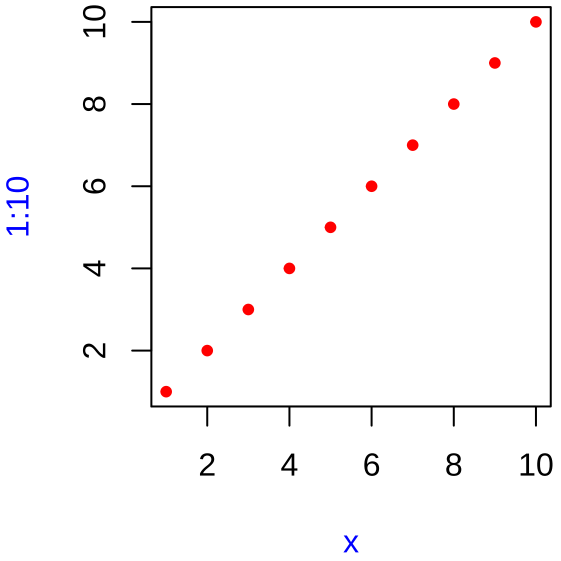
```
[1] 7
```

```
mean(1, 2, 3, na.omit = TRUE)
```

```
[1] 1
```

2. In the following call, explain how to find the documentation for the named arguments in the following function call:

```r
plot(1:10, col = "red", pch = 20, xlab = "x", col.lab = "blue")
```



3. Why does `plot(1:10, col = "red")` only colour the points, not the axes or labels? Read the source code of `plot.default()` to find out.

5.6 Exiting a function

Most functions exit in one of two ways[7]: either returning a value, indicating successful completion, or throwing an error, indicating failure. This section describes return values (implicit vs. explicit; visible vs. invisible), briefly discusses errors, and introduces exit handlers, which allow you to run code when a function exits, regardless of how it exits.

5.6.1 Implicit vs. explict returns

There are two ways that a function can return a value:

- Implicitly, where the last evaluated expression becomes the return value:

[7] Functions can exit in other more esoteric ways like signalling a condition that is caught by an exiting handler, invoking a restart, or pressing "Q" in an interactive browser.

```
j01 <- function(x) {
  if (x < 10) {
    0
  } else {
    10
  }
}
f(5)
```

```
[1] 52
```

```
f(15)
```

```
[1] 452
```

- Explicitly, by calling `return()`:

```
j02 <- function(x) {
  if (x < 10) {
    return(0)
  } else {
    return(10)
  }
}
```

5.6.2 Invisible values

Most functions return visibly: calling the function in an interactive context causes the result to be automatically printed.

```
j03 <- function() 1
j03()
```

```
[1] 1
```

However, it's also possible to return an `invisible()` value, which is not automatically printed.

```
j04 <- function() invisible(1)
j04()
```

You can verify that the value exists either by explicitly printing it or by wrapping in parentheses:

```
print(j04())
```

```
[1] 1
```

```
(j04())
```

```
[1] 1
```

Alternatively, use `withVisible()` to return the value and a visibility flag:

```
str(withVisible(j04()))
```

```
List of 2
 $ value  : num 1
 $ visible: logi FALSE
```

The most common function that returns invisibly is `<-`:

```
a <- 2
(a <- 2)
```

```
[1] 2
```

And this is what makes it possible to chain assignment:

```
a <- b <- c <- d <- 2
```

In general, any function called primarily for its side effects (like `<-`, `print()`, or `plot()`) should return an invisible value (typically the value of the first argument).

5.6.3 Errors

If a function can not complete its assigned task, it should throw an error with `stop()`, which immediately terminates the execution of the function.

```
j05 <- function() {
  stop("I'm an error")
  return(10)
}
j05()
```

```
Error in j05():
  I'm an error
```

Errors indicate that something has gone wrong, and force the user to handle them. Some languages (like C, go, and rust) rely on special return values to indicate problems, but in R you should always throw an error. You'll learn more about errors, and how to handle them, in [Conditions].

5.6.4 Exit handlers

Sometimes a function needs to make a temporary change to global state and you want to ensure those changes are restored when the function completes. It's painful to make sure you cleanup before any explicit return, and what happens if there's an error? Instead, you can set up an **exiting handler** that is called when the function terminates, regardless of whether it returns a value or throws an error.

To setup an exiting handler, call `on.exit()` with the code to be run. It will execute when the function exits, regardless of what causes it to exit:

```r
j06 <- function(x) {
  cat("Hello\n")
  on.exit(cat("Goodbye!\n"), add = TRUE)

  if (x) {
    return(10)
  } else {
    stop("Error")
  }
}

f(TRUE)
```

```
[1] 4
```

```r
f(FALSE)
```

```
[1] 2
```

Always set `add = TRUE` when using `on.exit()`. If you don't, each call to `on.exit()` will overwrite the previous exiting handler. Even when only registering a single handler, it's good practice to set `add = TRUE` so that you don't get an unpleasant surprise if you later add more exit handlers

`on.exit()` is important because it allows you to place clean-up actions next to actions with their cleanup operations.

```r
cleanup <- function(dir, code) {
  old_dir <- setwd(dir)
  on.exit(setwd(old), add = TRUE)

  old_opt <- options(stringsAsFactors = FALSE)
  on.exit(options(old_opt), add = TRUE)
}
```

When coupled with lazy evaluation, this leads to a very useful pattern for running a block of code in an altered environment:

```r
with_dir <- function(dir, code) {
  old <- setwd(dir)
  on.exit(setwd(old), add = TRUE)

  force(code)
}

getwd()
```

```
[1] "/home/linchao/rstudio/mybook"
```

```r
with_dir("~", getwd())
```

```
[1] "/home/linchao"
```

See the withr package for a collection of functions of this nature.

In R 3.4 and prior, `on.exit()` expressions are always run in the order in which they are created:

```r
f <- function() {
  on.exit(message("a"), add = TRUE)
  on.exit(message("b"), add = TRUE)
}
f()
```

```
a
```

```
b
```

This can make cleanup a little tricky if some actions need to happen in a specific order; typically you want the most recent added expression to be run first. In R 3.5 and later, you can control this by setting `after = FALSE`:

```
f <- function() {
  on.exit(message("a"), add = TRUE, after = FALSE)
  on.exit(message("b"), add = TRUE, after = FALSE)
}
f()
```

```
b
```

```
a
```

5.6.5 Exercises

1. What does `load()` return? Why don't you normally see these values?

2. What does `write.table()` return? What would be more useful?

3. How does the `chdir` parameter of `source()` compare to `in_dir()`? Why might you prefer one approach to the other?

4. Write a function that opens a graphics device, runs the supplied code, and closes the graphics device (always, regardless of whether or not the plotting code worked).

5. We can use `on.exit()` to implement a simple version of `capture.output()`.

   ```
   capture.output2 <- function(code) {
     temp <- tempfile()
     on.exit(file.remove(temp), add = TRUE, after = TRUE)

     sink(temp)
     on.exit(sink(), add = TRUE, after = TRUE)

     force(code)
     readLines(temp)
   }
   capture.output2(cat("a", "b", "c", sep = "\n"))
   ```

   ```
   [1] "a" "b" "c"
   ```

   Compare `capture.output()` to `capture.output2()`. How do the functions differ? What features have I removed to make the key ideas easier to see? How have I rewritten the key ideas to be easier to understand?

5.7 Function forms

> ”To understand computations in R, two slogans are helpful:
>
> - Everything that exists is an object.
> - Everything that happens is a function call.”
>
> — John Chambers

While everything that happens in R is a result of a function call, not all calls look the same. Function calls come in four varieties:

- In **prefix** form, the function name comes before its arguments, like `foofy(a, b, c)`. These constitute of the majority of function calls in R.
- In **infix** form, the function name comes inbetween its arguments, like `x + y`. Infix forms are used for many mathematical operators, as well as user-defined functions that begin and end with `%`.
- A **replacement** function assigns into what looks like a prefix function, like `names(df) <- c("a", "b", "c")`.
- **Special forms** like `[[`, `if`, and `for`, don't have a consistent structure and provide some of the most important syntax in R.

While four forms exist, you only need to use one, because any call can be written in prefix form. I'll demonstrate this property, and then you'll learn about each of the forms in turn.

5.7.1 Rewriting to prefix form

}}

An interesting property of R is every infix, replacement, or special form can be rewritten in prefix form. Rewriting in prefix form is useful because it helps you better understand the structure of the language, and it gives you the real name of every function. Knowing the real name of non-prefix functions is useful because it allows you to modify them for fun and profit.

The following example shows three pairs of equivalent calls, rewriting an infix form, replacement form, and a special form into prefix form.

```
x + y
`+`(x, y)

names(df) <- c("x", "y", "z")
```

```
`names<-`(df, c("x", "y", "z"))

for(i in 1:10) print(i)
`for`(i, 1:10, print(i))
```

Knowing the function name of a non-prefix function allows you to override its behaviour. For example, if you're ever feeling particularly evil, run the following code while a friend is away from their computer. It will introduce a fun bug: 10% of the time, 1 will be added to any numeric calculation inside of parentheses.

```
`(` <- function(e1) {
  if (is.numeric(e1) && runif(1) < 0.1) {
    e1 + 1
  } else {
    e1
  }
}
replicate(50, (1 + 2))
```

```
 [1] 3 3 4 3 3 3 3 3 3 3 4 3 3 3 3 4 3 3 3 3 3 3 3 3 4 3 3 3 4 3 3 3
[33] 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3
```

```
rm("(")
```

Of course, overriding built-in functions like this is a bad idea, but, as you'll learn about in [metaprogramming], it's possible to apply it only to selected code blocks. This provides a clean and elegant approach to writing domain specific languages and translators to other languages.

A more useful technique is to use this knowledge when using functional programming tools. For example, you could use `sapply()` to add 3 to every element of a list by first defining a function `add()`, like this:

```
add <- function(x, y) x + y
sapply(1:10, add, 3)
```

```
 [1]  4  5  6  7  8  9 10 11 12 13
```

But we can also get the same effect more simply by relying on the existing `+` function:

```
sapply(1:5, `+`, 3)
```

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

We'll explore this idea in detail in [functionals].

5.7.2 Prefix form {prefix-form}

The prefix form is the most common form in R code, and indeed in the majority of programming languages. Prefix calls in R are a little special because you can specify arguments in three ways:

- By position, like `help(mean)`.
- Using partial matching, like `help(to = mean)`.
- By name, like `help(topic = mean)`.

As illustrated by the following chunk, arguments are matched by exact name, then with unique prefixes, and finally by position.

```
k01 <- function(abcdef, bcde1, bcde2) {
  list(a = abcdef, b1 = bcde1, b2 = bcde2)
}
str(k01(1, 2, 3))
```

```
List of 3
 $ a : num 1
 $ b1: num 2
 $ b2: num 3
```

```
str(k01(2, 3, abcdef = 1))
```

```
List of 3
 $ a : num 1
 $ b1: num 2
 $ b2: num 3
```

```
# Can abbreviate long argument names:
str(k01(2, 3, a = 1))
```

```
List of 3
 $ a : num 1
 $ b1: num 2
 $ b2: num 3
```

```
# But this doesn't work because abbreviation is ambiguous
str(k01(1, 3, b = 1))
```

```
Error in k01(1, 3, b = 1):
  引數 3 有多個與之相對應的正式引數
```

Generally, only use positional matching for the first one or two arguments; they will be the most commonly used, and most readers will know what they are. Avoid using positional matching for less commonly used arguments, and never use partial matching. See the tidyverse style guide, http://style.tidyverse.org/syntax.html#argument-names, for more advice.

5.7.3 Infix functions

Infix functions are so called because the function name comes **in**between its arguments, and hence infix functions have two arguments. R comes with a number of built-in infix operators: `:`, `::`, `:::`, `$`, `@`, `^`, `*`, `/`, `+`, `-`, `>`, `>=`, `<`, `<=`, `==`, `!=`, `!`, `&`, `&&`, `|`, `||`, `~`, `<-`, and `<<-`. You can also create your own infix functions that start and end with `%`, and base R uses this to additionally define `%%`, `%*%`, `%/%`, `%in%`, `%o%`, and `%x%`.

Defining your own infix function is simple. You create a two argument function and bind it to a name that starts and ends with `%`:

```
`%+%` <- function(a, b) paste0(a, b)
"new " %+% "string"
```

```
[1] "new string"
```

The names of infix functions are more flexible than regular R functions: they can contain any sequence of characters except "%". You will need to escape any special characters in the string used to define the function, but not when you call it:

```
`% %` <- function(a, b) paste(a, b)
`%/\\%` <- function(a, b) paste(a, b)

"a" % % "b"
```

```
[1] "a b"
```

```
"a" %/\% "b"
```

```
[1] "a b"
```

R's default precedence rules mean that infix operators are composed from left to right:

```
`%-%` <- function(a, b) paste0("(", a, " %-% ", b, ")")
"a" %-% "b" %-% "c"
```

```
[1] "((a %-% b) %-% c)"
```

There are two special infix functions that can be called with a single argument: `+` and `-`.

```
-1
```

```
[1] -1
```

```
+10
```

```
[1] 10
```

5.7.4 Replacement functions

Replacement functions act like they modify their arguments in place, and have the special name `xxx<-`. They must have arguments named `x` and `value`, and must return the modified object. For example, the following function allows you to modify the second element of a vector:

```
`second<-` <- function(x, value) {
  x[2] <- value
  x
}
```

Replacement functions are used by placing the function call on the LHS of `<-`:

```
x <- 1:10
second(x) <- 5L
x
```

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

I say they "act" like they modify their arguments in place, because, as discussed in [Modify-in-place], they actually create a modified copy. We can see that by using `tracemem()`:

```
x <- 1:10
tracemem(x)
#> <0x7ffae71bd880>

second(x) <- 6L
#> tracemem[0x7ffae71bd880 -> 0x7ffae61b5480]:
#> tracemem[0x7ffae61b5480 -> 0x7ffae73f0408]: second<-
```

If you want to supply additional arguments, they go inbetween `x` and `value`:

```
`modify<-` <- function(x, position, value) {
  x[position] <- value
  x
}
modify(x, 1) <- 10
x
```

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

When you write `modify(x, 1) <- 10`, behind the scenes R turns it into:

```
x <- `modify<-`(x, 1, 10)
```

Combining replacement with other functions requires more complex translation. For example, this:

```
x <- c(a = 1, b = 2, c = 3)
names(x)
```

```
[1] "a" "b" "c"
```

```
names(x)[2] <- "two"
names(x)
```

```
[1] "a"   "two" "c"
```

Is translated into:

```
`*tmp*` <- x
x <- `names<-`(`*tmp*`, `[<-`(names(`*tmp*`), 2, "two"))
rm(`*tmp*`)
```

(Yes, it really does create a local variable named *tmp*, which is removed afterwards.)

5.7.5 Special forms

Finally, there are a bunch of language features that are usually written in special ways, but also have prefix forms. These include parentheses:

- `(x)` (`` `(`(x) ``)
- `{x}` (`` `{`(x) ``).

The subsetting operators:

- `x[i]` (`` `[`(x, i) ``)
- `x[[i]]` (`` `[[`(x, i) ``)

And the tools of control flow:

- `if (cond) true` (`` `if`(cond, true) ``)
- `if (cond) true else false` (`` `if`(cond, true, false) ``)
- `for(var in seq) action` (`` `for`(var, seq, action) ``)
- `while(cond) action` (`` `while`(cond, action) ``)
- `repeat expr` (`` `repeat`(expr) ``)
- `next` (`` `next`() ``)
- `break` (`` `break`() ``)

Finally, the most complex is the "function" function:

- `function(arg1, arg2) {body}` (`` `function`(alist(arg1, arg2), body, env) ``)

Knowing the name of the function that underlies the special form is useful for getting documentation. `?(` is a syntax error; `` ?`(` `` will give you the documentation for parentheses.

Note that all special forms are implemented as primitive functions (i.e. in C); that means printing these functions is not informative:

```
`for`
```

```
.Primitive("for")
```

5.8 Invoking a function

Suppose you had a list of function arguments:

```
args <- list(1:10, na.rm = TRUE)
```

How could you then send that list to `mean()`? In base R, you need `do.call()`:

```
do.call(mean, args)
```

```
[1] 5.5
```

```
# Equivalent to
mean(1:10, na.rm = TRUE)
```

```
[1] 5.5
```

5.8.1 Exercises

1. Rewrite the following code snippets into prefix form:

   ```
   1 + 2 + 3

   1 + (2 + 3)

   if (length(x) <= 5) x[[5]] else x[[n]]
   ```

2. Clarify the following list of odd function calls:

   ```
   x <- sample(replace = TRUE, 20, x = c(1:10, NA))
   y <- runif(min = 0, max = 1, 20)
   cor(m = "k", y = y, u = "p", x = x)
   ```

3. Explain why the following code fails:

```
modify(get("x"), 1) <- 10
#> Error: target of assignment expands to non-language object
```

4. Create a replacement function that modifies a random location in a vector.

5. Write your own version of `+` that will paste its inputs together if they are character vectors but behaves as usual otherwise. In other words, make this code work:

   ```
   1 + 2
   #> [1] 3

   "a" + "b"
   #> [1] "ab"
   ```

6. Create a list of all the replacement functions found in the base package. Which ones are primitive functions? (Hint use `apropros()`)

7. What are valid names for user-created infix functions?

8. Create an infix `xor()` operator.

9. Create infix versions of the set functions `intersect()`, `union()`, and `setdiff()`. You might call them `%n%`, `%u%`, and `%/%` to match conventions from mathematics.

5.9 Quiz answers

1. The three components of a function are its body, arguments, and environment.

2. `f1(1)()` returns 11.

3. You'd normally write it in infix style: `1 + (2 * 3)`.

4. Rewriting the call to `mean(c(1:10, NA), na.rm = TRUE)` is easier to understand.

5. No, it does not throw an error because the second argument is never used so it's never evaluated.

6. See infix and replacement functions.

7. You use `on.exit()`; see on exit for details.

Chapter 6

資料探索

6.1 常用函數

- Mean
- Median
- Quartile
- Percentile
- Range
- Interquartile Range
- Box Plot
- Variance
- Standard Deviation
- Covariance
- Correlation Coefficient
- Central Moment
- Skewness
- Kurtosis

```
mean(mtcars$mpg)
```

```
[1] 20.1
```

```
median(mtcars$mpg)
```

```
[1] 19.2
```

```
min(mtcars$mpg)
```

```
[1] 10.4
```

```
max(mtcars$mpg)
```

```
[1] 33.9
```

```
quantile(mtcars$mpg)
```

```
  0%  25%  50%  75% 100% 
10.4 15.4 19.2 22.8 33.9 
```

```
quantile(mtcars$mpg,c(0.2))
```

```
 20% 
15.2 
```

```
require(moments)
skewness(mtcars$mpg)
```

```
[1] 0.64
```

```
kurtosis(mtcars$mpg)
```

```
[1] 2.8
```

## 匯入其他格式的資料

以 `stata` 為例:

6.2 來自 wooldriidge 的資料集

base cheat sheet

```
load('resources/affairs.RData')
# 原始資料 kid 是整數，加入標籤成為 factor
haskids <- factor(affairs$kids,labels=c("no","yes"))
mlab <- c("very unhappy","unhappy","average","happy", "very happy")
marriage <- factor(affairs$ratemarr, labels=mlab)

# Frequencies for having kids:
table(haskids)
```

```
haskids
 no yes 
171 430 
```

```
# Marriage ratings (share):
prop.table(table(marriage))
```

```
marriage
very unhappy      unhappy      average        happy   very happy 
      0.0266       0.1098       0.1547       0.3228       0.3860 
```

```
# Contigency table: counts (display & store in var.)
(countstab <- table(marriage,haskids))
```

```
              haskids
marriage        no yes
  very unhappy   3  13
  unhappy        8  58
  average       24  69
  happy         40 154
  very happy    96 136
```

```
# Share within "marriage" (i.e. within a row):
prop.table(countstab, margin=1)
```

```
              haskids
marriage          no   yes
  very unhappy 0.188 0.812
  unhappy      0.121 0.879
  average      0.258 0.742
  happy        0.206 0.794
  very happy   0.414 0.586
```

```
# Share within "haskids"  (i.e. within a column):
prop.table(countstab, margin=2)
```

```
              haskids
marriage           no    yes
  very unhappy 0.0175 0.0302
  unhappy      0.0468 0.1349
  average      0.1404 0.1605
  happy        0.2339 0.3581
  very happy   0.5614 0.3163
```

如果不知道函數 prop.table()

看一下資料結構

```
str(x<-table(haskids))
```

```
 'table' int [1:2(1d)] 171 430
 - attr(*, "dimnames")=List of 1
  ..$ haskids: chr [1:2] "no" "yes"
```

上面的結構顯示,

- int [1:2(1d)] 171 430：整數向量，長度由 1 到 2，是個 1 維向量。元素 1 到 2 的內容分別是 171,430。
- 可以由 dimnames(x) 得到一個 list, 裡面只有一個元素，這個元素又是一個 char list(長度為 2)。
  因此可以試試看

```
x/sum(x)
```

```
haskids
   no   yes
0.285 0.715
```

結果應該和 prop.table() 一樣。

那麼 2 way 呢？ 也是先看結構:

```
str(countstab <- table(marriage,haskids))
```

```
 'table' int [1:5, 1:2] 3 8 24 40 96 13 58 69 154 136
 - attr(*, "dimnames")=List of 2
  ..$ marriage: chr [1:5] "very unhappy" "unhappy" "average" "happy" ...
  ..$ haskids : chr [1:2] "no" "yes"
```

```
 countstab/rowSums(countstab)
```

```
              haskids
marriage          no   yes
  very unhappy 0.188 0.812
  unhappy      0.121 0.879
  average      0.258 0.742
  happy        0.206 0.794
  very happy   0.414 0.586
```

```
countstab/colSums(countstab)
```

```
              haskids
marriage            no    yes
  very unhappy 0.0175 0.0302
  unhappy      0.0186 0.3392
  average      0.1404 0.1605
  happy        0.0930 0.9006
  very happy   0.5614 0.3163
```

```
#??countstab/colsum(countstab)
```

```
load('resources/ceosal1.RData')
# sample average:
mean(ceosal1$salary)
```

```
[1] 1281
```

```
# sample median:
median(ceosal1$salary)
```

```
[1] 1039
```

```
#standard deviation:
sd(ceosal1$salary)
```

```
[1] 1372
```

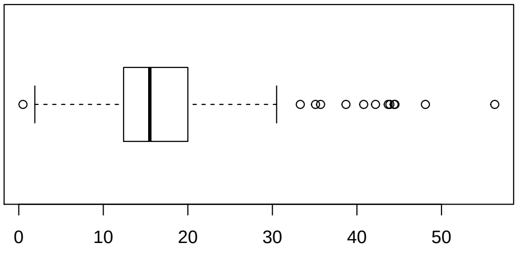
```
# summary information:
summary(ceosal1$salary)
```

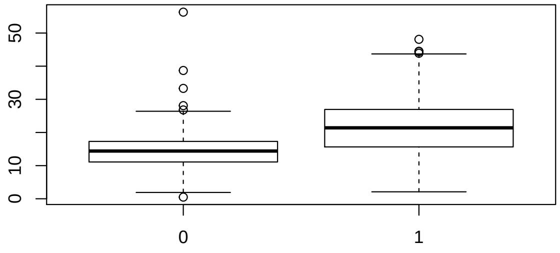
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    223     736    1039    1281    1407   14822 
```

```
# correlation with ROE:
cor(ceosal1$salary, ceosal1$roe)
```

```
[1] 0.115
```

```
boxplot(ceosal1$roe,horizontal = T)
boxplot(ceosal1$roe~ceosal1$consprod)
```





6.3 t-test

6.3.1 Manually enter raw data from Wooldridge, Table C.3:

```
SR87<-c(10,1,6,.45,1.25,1.3,1.06,3,8.18,1.67,.98,1,.45,
                                     5.03,8,9,18,.28,7,3.97)
SR88<-c(3,1,5,.5,1.54,1.5,.8,2,.67,1.17,.51,.5,.61,6.7,
                                           4,7,19,.2,5,3.83)
```

6.3.2 Calculate Change (the parentheses just display the results):

```
(Change <- SR88 - SR87)
```

```
 [1] -7.00  0.00 -1.00  0.05  0.29  0.20 -0.26 -1.00 -7.51 -0.50
[11] -0.47 -0.50  0.16  1.67 -4.00 -2.00  1.00 -0.08 -2.00 -0.14
```

6.3.3 Ingredients to CI formula

```
(avgCh<- mean(Change))
```

```
[1] -1.15
```

```
(n    <- length(Change))
```

```
[1] 20
```

```
(sdCh <- sd(Change))
```

```
[1] 2.4
```

```
(se   <- sdCh/sqrt(n))
```

```
[1] 0.537
```

```
(c    <- qt(.975, n-1))
```

```
[1] 2.09
```

6.3.4 Confidence intervall:

```
c( avgCh - c*se, avgCh + c*se )
```

```
[1] -2.278 -0.031
```

驗證

```
v.n <- length(Change)
v.mean <- sum(Change)/v.n
v.std <-sqrt(sum((Change-v.mean)^2)/(v.n-1))
v.se <- v.std/sqrt(v.n)
rst<-list(v.n,v.mean,v.std,v.se)
rst
```

```
[[1]]
[1] 20

[[2]]
[1] -1.15

[[3]]
[1] 2.4

[[4]]
[1] 0.537
```

練習：黑白種，用同一個 CV 找工作。

```
load('resources/audit.dta')
head(audit)
```

```
  w b y
1 1 1 0
2 1 1 0
3 1 1 0
4 1 1 0
5 1 1 0
6 0 0 0
```

y 是前兩個欄位相減。以第 1 筆資料為例，兩個都是 1，沒差別 (y=0)。

6.3.5 by hand

t statistic for H0: mu=0:

```
(t <- avgy/se)
```

```
[1] -4.28
```

自由度 (d.f.) 為 n-1 =240 的 t 分配，其臨界值 Critical values(如下):

```
alpha.one.tailed = c(0.1, 0.05, 0.025, 0.01, 0.005, .001)
CV <- qt(1 - alpha.one.tailed, n-1)
cbind(alpha.one.tailed, CV)
```

```
     alpha.one.tailed   CV
[1,]            0.100 1.29
[2,]            0.050 1.65
[3,]            0.025 1.97
[4,]            0.010 2.34
[5,]            0.005 2.60
[6,]            0.001 3.12
```

6.4 by 函數 t.test()

H0: y=5，H1:y>5，信賴區間 0.99 t.test(y，mu=5，al.ternative=“greater”，conf.l.evel.=0.99) 利用函數算:

```
# ex.C.3: two-sided CI
t.test(audit$y)
# ex.C.5 & C.7: 1-sided test:
t.test(audit$y, alternative="less")
```

```
One Sample t-test

data:  audit$y
t = -4.2768, df = 240, p-value = 2.739e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1939385 -0.0716217  **說明：mu +/- c.level*(se) 見手算例   **
sample estimates:
 mean of x
-0.1327801


    One Sample t-test

data:  audit$y
t = -4.2768, df = 240, p-value = 1.369e-05
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
        -Inf -0.08151529
```

```
sample estimates:
 mean of x
-0.1327801
```

6.5 分組

6.5.1 方法 1: 利用 for 迴圈

大綱是：先探索一下我們可能用到的 R 語言技巧，然後是整個程式：

```
age    = c(23.0883, 25.8344, 29.4648, 32.7858, 33.6372,
           34.935,  35.2115, 35.2115,  5.2115, 36.7803)
group  = c(1, 1, 1, 2, 1, 1, 2, 2, 2, 1)
dframe = data.frame(age=age, group=group)
summary(dframe)
```

```
      age           group    
 Min.   : 5.2   Min.   :1.0  
 1st Qu.:26.7   1st Qu.:1.0  
 Median :33.2   Median :1.0  
 Mean   :29.2   Mean   :1.4  
 3rd Qu.:35.1   3rd Qu.:2.0  
 Max.   :36.8   Max.   :2.0  
```

函數 summary()

```
aggregate(dframe$age, by=list(dframe$group), FUN=mean)[2]
```

```
     x
1 30.6
2 27.1
```

data.frame 的 record 篩選和 matrix 的 row 篩選不一樣，後者是 row based, 前者是 column base 因此，篩選的方法如下：

```
ft<-factor(dframe$group)
gn<-length(levels(ft)) # nlevels(ft)

gf<-ft==levels(ft)[1]
g1<-dframe[gf,]
g1
```

```
    age group
1  23.1     1
2  25.8     1
3  29.5     1
5  33.6     1
6  34.9     1
10 36.8     1
```

除了上面的篩選方式以外，也可以

```
ft<-factor(dframe$group)
gn<-length(levels(ft)) # nlevels(ft)

gf<-as.numeric(levels(ft)[1])
g1<-dframe[group %in% gf,] # 不是 gf %in% group
rst<-summary(g1)
rst[4,1]
```

```
[1] "Mean   :30.6  "
```

```
rst[4,2]
```

```
[1] "Mean   :1  "
```

最後的程式碼：

```
ft<-factor(dframe$group)
rst <- as.list(numeric(nlevels(ft)))
names(rst)<-levels(ft)
for (gi in levels(ft) ){
  gf<-as.numeric(gi)
  g1<-dframe[group %in% gf,]
  t<-summary(g1)
  rst[gi]=t[4,1]


}
print(rst)
```

```
$`1`
[1] "Mean   :30.6  "

$`2`
[1] "Mean   :27.1  "
```

函數 `summary()` 的結果是一個 table，而且裡面的元素是字串，如果我們只要「數字」平均，考慮使用字串函數 `sub()` as.numeric(sub(’.*:‘,”, summary(dframe)[4,1]))

6.5.2 方法 2: tapply

```
f1<-function(im)
{
  c(min(im),median(im),mean(im),max(im))
}
rst<-tapply(dframe$age,dframe$group,FUN = f1)
rstm<-unlist(rst) # unlist 會把 rst 變成 1 維向量
rstm<-matrix(rstm,ncol=2) # hard code2 , try `length(levels(factor(dframe$group)))`
rownames(rstm)<-c("min","median","mean","max")
colnames(rstm)<-paste(" 組別:",levels(factor(dframe$group)))
rstm
```

```
       組別: 1 組別: 2
min       23.1    5.21
median    31.6   34.00
mean      30.6   27.11
max       36.8   35.21
```

6.5.2.1 方法 3 利用其他套件

```
library(dplyr)
dt <- data.frame(age=rchisq(20,10), group=sample(1:2,20, rep=T))
grp <- group_by(dt, group)
summarise(grp, mean=mean(age), sd=sd(age))
detach("package:dplyr", unload=TRUE)
```

6.6 plot

這裡的分組不能用之前的 hard code 中提到的方法，因為各組可能數量不同。

```
v1<-c(1,2,3)
v2<-c(4,4)
cbind(v1,v2)
```

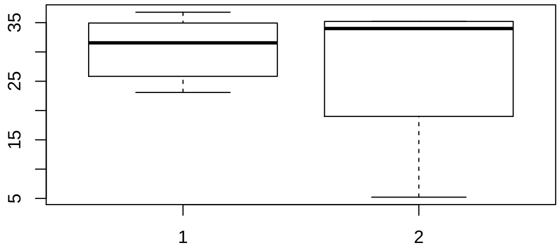
```
Warning in cbind(v1, v2): number of rows of result is not a multiple
of vector length (arg 2)
```

```
     v1 v2
[1,]  1  4
[2,]  2  4
[3,]  3  4
```

```
v1<-c(1,2,3)
v2<-c(4,4)
n<-max(sapply(list(v1,v2),length))
length(v1)<-n
length(v2)<-n
cbind(v1,v2)
```

```
     v1 v2
[1,]  1  4
[2,]  2  4
[3,]  3 NA
```

```
age    = c(23.0883, 25.8344, 29.4648, 32.7858, 33.6372,
           34.935,  35.2115, 35.2115,  5.2115, 36.7803)
group  = c(1, 1, 1, 2, 1, 1, 2, 2, 2, 1)
dframe = data.frame(age=age, group=group)
rst<-tapply(age,group,FUN = matrix,ncol=1) # 每個組成為一個 list 元素
nrst<-sapply(rst,length) #<2> 每一個組的長度
n<-max(nrst)  # 最大長度
mrst<-sapply(rst,`length<-`,n) # 重設每組長度<1>
boxplot(mrst)
```



<1> FUN.Name 會跑到 mathch.fun() 這個函數，而這個函數是用字串搜尋

m2: 現在想要把 rst 轉成 data.frame。
先看一下示範，如何把 list 轉成 data.frame

```
# 不是這種
test1 <- list( c(a='a',b='b',c='c'),
               c(a='d',b='e',c='f'))
as.data.frame(test1)
```

```
  c.a....a...b....b...c....c.. c.a....d...b....e...c....f..
a                            a                            d
b                            b                            e
c                            c                            f
```

```
# ok
test2 <- list(a = c(1, 2, 3), b = c(4, 5, 6))
as.data.frame(test2)
```

```
  a b
1 1 4
2 2 5
3 3 6
```

Chapter 7

R Packages

7.1 基本

7.1.1 套件在哪裡

- R package and Github

  note: 可以 DOS 指令 tree . 列出目錄樹結構

```
R.home()
```

```
[1] "/usr/lib/R"
```

```
system.file()
```

```
[1] "/usr/lib/R/library/base"
```

7.1.2 搜尋套件中的檔案

system.file(“help”, “AnIndex”, package = “splines”)

結果:[1] “C:/PROGRA$_{1/\text{MICROS}}$4/RCLIEN~1/R_SERVER/library/splines/help/AnIndex”
解釋：在套件 splines 中的根目錄中搜尋子目錄 help，找出名稱為 AnIndex 的檔案，
並傳回路徑名稱。

練習：
例如, 套件 epuRate 的安裝目錄為 `D:\RSTUDIO\RMD\RPACK\EPURATE-MASTER\INST`,
目錄結構如下：

```
D:\RSTUDIO\RMD\RPACK\EPURATE-MASTER\INST
 rmarkdown
     templates
         basic
            skeleton
         epurate
            resources
            skeleton
         PCTG
              resources
              skeleton
```

參考上面的目錄結構，則下面程式執行的結果 css,header, template 分別是什麼?

```
  css <- system.file("rmarkdown", "templates", "epurate" ,"resources", "style.css", pa
  header <- system.file("rmarkdown", "templates", "epurate" ,"resources", "header.html"
  template <- system.file("rmarkdown", "templates", "epurate" ,"resources", "template_
```

hint:

```
css:  D:\RSTUDIO\RMD\RPACK\EPURATE-MASTER\INST\rmarkdown\templates\eurate\resouces\styl
```

問題:

```
system.file('rmarkdown')
```

```
[1] ""
```

為什麼傳回空值: ”” ?
hint 沒有指定套件名稱，因此預設為 `base`，而 base 的根目錄中，找不到子目錄或檔案為 `rmarkdown` 的檔案。

下面驗證驗證這句話「找不到子目錄或檔案為 `rmarkdown` 的檔案」

```
list.dirs(system.file())
```

```
[1] "/usr/lib/R/library/base"      "/usr/lib/R/library/base/demo"
[3] "/usr/lib/R/library/base/help" "/usr/lib/R/library/base/html"
[5] "/usr/lib/R/library/base/Meta" "/usr/lib/R/library/base/R"
```

```
list.files(system.file())
```

```
[1] "CITATION"    "demo"        "DESCRIPTION" "help"
[5] "html"        "INDEX"       "Meta"        "R"
```

> note: list.files 傳回的不僅是檔案名稱，還有子目錄名稱。

`system.file()` 傳回套件 `base` 的根目錄
`list.dirs(system.file())` 傳回上述根目錄的所有子目錄。

驗證 `system.file()` 的搜尋範圍包括子目錄和檔案名稱。下面的參數，一個是子目錄 html 一個是檔案 INDEX，都傳回路徑名稱。

```
system.file("html")
```

```
[1] "/usr/lib/R/library/base/html"
```

```
system.file("INDEX")
```

```
[1] "/usr/lib/R/library/base/INDEX"
```

7.1.3 哪個函數屬於哪個套件?

可以直接打入函數名稱但是不要有括號，看最後一行:

```
R.home <-
function (component = "home")
{
    rh <- .Internal(R.home())
    switch(component, home = rh, bin = if (.Platform$OS.type ==
        "windows" && nzchar(p <- .Platform$r_arch)) file.path(rh,
        component, p) else file.path(rh, component), share = if (nzchar(p <- Sys.getenv("R_SHARE_
        component), doc = if (nzchar(p <- Sys.getenv("R_DOC_DIR"))) p else file.path(rh,
        component), include = if (nzchar(p <- Sys.getenv("R_INCLUDE_DIR"))) p else file.path(rh,
        component), modules = if (nzchar(p <- .Platform$r_arch)) file.path(rh,
        component, p) else file.path(rh, component), file.path(rh,
        component))
}
<bytecode: 0x3183948>
<environment: namespace:base>
```

7.2 library() vs require()

7.3 測試 debug

- RTVS debug
- 互動 R document

y 是局部變數。

```
x <- 2
g <- function() {
    y <- 1
    c(x, y)
}
g()
```

```
[1] 2 1
```

```
y
```

```
Error in eval(expr, envir, enclos):
  找不到物件 'y'
```

```
rm(x, g)
```

7.4 自製 package

- R 包製作
- package example R packages book on documentation
  (Hilary Parker's is the classic), and if you're in the second camp you probably already know what you're doing (and if not, Hadley Wickham's R packages book is an excellent and thorough guide).

7.4.1 自製 package 範例

1. 本範例中 package 只有一個 data.frame, 一個函數.

2. 執行完 package.seleton() 以後, 必須在 man 子目錄中修改每個 RD 檔案,title 裡面必須有內容。

3. 然後才執行 build()

範例需要的檔案：trees91.csv

```
rm(list = ls())
ufc <- read.csv('./resource/trees91.csv')
vol.m3 <- function(dbh.cm, height.m, multiplier = 0.5) {
    vol.m3 <- pi * (dbh.cm / 200) ^ 2 * height.m * multiplier
     return(vol.m3)

}
package.skeleton(name = "xxx", path = "./packages", force = TRUE)

library(devtools)
build("./packages/xxx")
build("./packages/xxx", binary = TRUE)
```

7.5 Packing up your data

Rstudio, package `devtools` , RMD 常識。

7.6 What is a package?

最小的套件一般包括一個根目錄,下有一個子目錄 `R\`,一個檔案 `DESCRIPTION`(package 叫什麼名子，依賴於哪些其他套件).

當載入套件的時候，會執行套件中位於子目錄 `R/`中的`.R`。

7.6.1 也可以包含資料

資料一般位於子目錄 `data/` 中。副檔名一般是`.RData`，`.csv`，`.R`。但是不受限制。
有三個方法可以從套件存取資料。例如，

1. 利用`::`

```
ggplot2::diamonds %>% head()
```

2. `library()`

```
library(ggplot2)
diamonds %>% head()
```

3. data()

```
data("diamonds", package="ggplot2")
ls()
```

note: 2,3 的不同是，後者直接導入 global 空間。

7.7 怎樣自訂套件

7.7.1 利用 devtools 建立骨架 (package skeleton)

7.7.1.1 With (the slightly less easy way)

1. 建立套件骨架 devtools::create('~/mypackage').
2. 編輯 DESCRIPTION.
3. 程式碼放入子目錄 R/.

7.7.2 工作流程

7.7.2.1 測試

如果目前的工作目錄就是套件根目錄，那麼加入資料和寫完程式碼以後，就可以使用指令 devtools::load_all()，或是利用 Rstudio 的按鍵。可以參考這裡。

7.7.2.2 Installed package workflow

如果套件已經安裝，並且被其他的專案中使用，那麼程式碼變動的話，就必須 3 個步驟

1. Edit code (or add data).
2. Re-generate documentation and namespace: devtools::document('/path/to/pkg')
3. Install: devtools::install('/path/to/pkg').
4. Reload: devtools::reload(inst('pkg')).

7.7.3 Documentation and exports

當套建立用指令 library() 載入時，預設上，所有的變數和函數都是 **private** 並不會被加入到 global 環境。而檔案 NAMESPACE 則是用來指定要 ___public___ 的函數。如果不想要手動編輯檔案 NAMESPACE，則可以使用特殊指令，產生如下的格式：

```
#' Short description of what this does
#'
#' Longer description of what this does. Approximately a paragraph.
#'
#' @param x The first thing
#' @param y The second thing.
#' @return The thing that comes out of this function
#'
#' @export (do export this in NAMESPACE)
a_function <- function(x,y) {
  return x+y
}
```

然後利用指令，devtools::document()，會自動更新 NAMESPACE 並建立函數相關的説明檔。例如，man/a_function.Rd. 這樣這個函數，就像其他一般的函數一樣，可以求助，也可以在指令 library() 之後呼叫調用。

如果只是自行測試，指令 devtools::load_all() 就可以了。但是通常要散發，也就是利用 devtools::install() 安裝套件。

7.8 資料放入套件

最簡單的方法，直接將資料檔（.RData, .csv, etc.）置於 data/子目錄。

比較不會出錯的方式是利用 usethis 套件（取代舊版 devtools ）。基本上用到兩個函數 userthis::use_data_raw() userthis::use_data()

1. usethis::use_data_raw()，會建立子目錄 data-raw/:
2. 在 data-raw/子目錄中，放進非 R 格式的資料，例如，csv 檔案。
3. 處理資料的程式碼，任意 Rscript 名稱，然後手動執行，（不會被自動執行）置於 data-raw/ ，如下：

   ```
   experiment1 <-
     read.csv('expt1.csv')
   usethis::use_data(experiment1)
   ```

   This saves data/experiment1.RData in your package directory (make sure you've setwd() to the package directory…)

4. 測試:

```
devtools::load_all()
experiment1 %>% head()
## or use data() to put it in the global environment
data("experiment1")
```

Chapter 8

Environments

參考 https://holtzy.github.io/Pimp-my-rmd/

8.1 Introduction

The environment is the data structure that powers scoping. 相關概念:lexical scoping, namespaces, and R6 classes。

這個文件需要

```
devtools::install_github("tidyverse/rlang")
```

Quiz

If you can answer the following questions correctly, you already know the most important topics in this chapter. You can find the answers at the end of the chapter in answers.

1. List at least three ways that an environment is different to a list.
2. What is the parent of the global environment? What is the only environment that doesn't have a parent?
3. What is the enclosing environment of a function? Why is it important?
4. How do you determine the environment from which a function was called?
5. How are `<-` and `<<-` different?

Outline

- Environment basics introduces you to the basic properties of an environment and shows you how to create your own.
- Recursing over environments provides a function template for computing with environments, illustrating the idea with a useful function.
- Explicit environments briefly discusses three places where environments are useful data structures for solving other problems.

Prerequisites

這個章節利用了套件 `rlang` 裡的函數，來探索環境物件。

在 `rlang` 套件中,`env_` 函數是設計用來和 pipe 一起工作的, 這裡不深入。

`global_env()` 和 `globalenv()` 的執行結果一樣。

```
.GlobalEnv
```

```
<environment: R_GlobalEnv>
```

```
globalenv()
```

```
<environment: R_GlobalEnv>
```

```
global_env()
```

```
<environment: R_GlobalEnv>
```

```
.BaseNamespaceEnv
```

```
<environment: namespace:base>
```

```
current_env() #
```

```
<environment: R_GlobalEnv>
```

8.2 Environment basics

基本上一個 environment 類似名稱串列 (named list), 但是有 4 個例外:

- 名稱唯一 (就是變數唯一)
- 名稱沒有順序關係
- 會有一個 parent
- 當改變的時候, 不會自動複製 (Environments are not copied when modified).

分別探索上面四點:

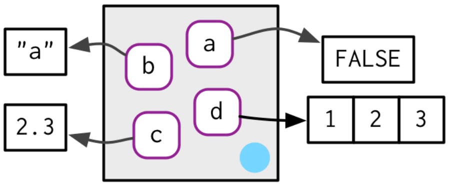
8.2.1 Basics

要建立 environment, 使用 `rlang::env()`. 類似使用 `list()`, 也是一組名稱-值的配對。:

```
e1 <- env(
  a = FALSE,
  b = "a",
  c = 2.3,
  d = 1:3
)
```

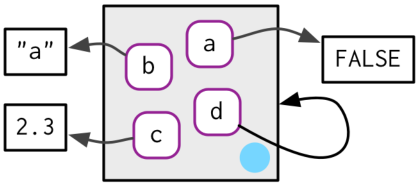
建立 environment 物件, 利用函數 `new.env()` 不用管參數 `hash` 和 `size`。注意不能利用 `$<-`同時定義和建立 parameters; 例如, e1 <- env(** a <- FALSE **) # error .

environment 物件可以想成是一個袋子, 或是 names 集合。因為沒有次序關係



就像在names and values, 討論的, 這個物件是參考為基礎.(in C concept) 不會有 copy on modifying。而且, 環境物件可以自己指向自己 (recursion)

```
e1$d <- e1
```



沒有指派的環境變數, 只會顯示記憶體位址:

```
e1
```

```
<environment: 0x55a96e85e780>
```

要知道內容可以使用 `env_print()` :

```
env_print(e1)
```

```
<environment: 0x55a96e85e780>
parent: <environment: global>
bindings:
 * a: <lgl>
 * b: <chr>
 * c: <dbl>
 * d: <env>
```

想要知道目前有哪些 binding(名稱-值配對) 可以利用 `env_names()`

```
env_names(e1)
```

```
[1] "a" "b" "c" "d"
```

要列出環境下的繫結，在 R 3.2.0 以上, 可以使用函數 `names()` , 之前的版本則是 `ls()`, 但是要注意的是 ls 的參數 `all.names` 內設是 `FALSE` 因此. 開頭的看不到。.

8.2.2 Important environments

另外參考Special environments。`current_env()` 可以知道目前程式碼的執行環境。例如, 當我們互動執行 RCODE 的時候, 環境通常是總體環境, 或者由函數 `global_env()` 可以得到。這個總體環境有時候就叫 “workspace”, 同時, 這也是函數外面所有互動計算發生的地方。環境物件的比較不能用 `==`, 只能用函數 `identical()`。

```
identical(global_env(), current_env())
```

```
[1] TRUE
```

```
global_env() == current_env()
```

```
Error in global_env() == current_env():
  只有基元或串列類型才能做比較 (1)
```

- `globalenv()` 和`.GlobalEnv`: 拿到 global environmentand。
- `environment()`: 拿到目前的環境

global environment 的名稱為 `R_GlobalEnv` 。

```
global_env()
```

```
<environment: R_GlobalEnv>
```

```
current_env()
```

```
<environment: R_GlobalEnv>
```

```
.GlobalEnv
```

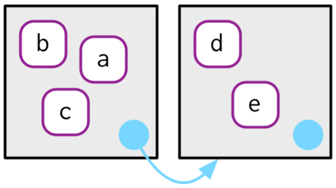
```
<environment: R_GlobalEnv>
```

8.2.3 Parents

每一個環境物件都有一個 *parent*。*parent* 也一個環境物件。在方塊圖中,parent 以藍色圈表示, 並用箭頭指向另一個環境物件。

這個 parent 用來建立 lexical scoping: 如果 name 沒有在某個環境物件找到,R 會重複的在 parent 中找。函數 `env()` 可以用來建立一個沒有名字的環境 You can set the parent environment by supplying an unnamed argument to `env()`. If you don't supply it, it defaults to the current environment.

```
e2a <- env(d = 4, e = 5)
e2b <- env(e2a, a = 1, b = 2, c = 3)
```



函數 `env_parent()` 可以用來找出某個環境物件的 parent:

```
env_parent(e2b)
```

```
<environment: 0x55a975b59410>
```

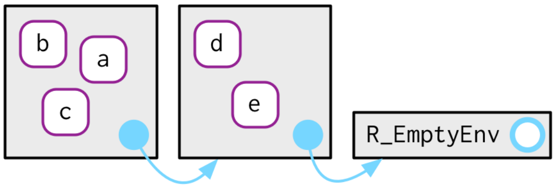
```
env_parent(e2a)
```

```
<environment: R_GlobalEnv>
```

parent.env() === env_parent()

所有的環境物件中只有一個名稱為 `R_EmptyEnv` 的物件沒有 parent(用空心藍色表示):

```
e2c <- env(empty_env(), d = 4, e = 5)
e2d <- env(e2c, a = 1, b = 2, c = 3)
```



emptyenv() === empty_env()

試圖利用函數 `env_parent()` 找空環境物件的 parent 會發生錯誤:

```
env_parent(empty_env())
```

```
Error: The empty environment has no parent
```

函數 `env_parents()` 可以找出目前環境物件的所有祖先: 這個函數會繼續直到遇上 global environment 或是空環境物件。上述過程可以利用 `last` 環境物件控制。

```
env_parents(e2b)
```

```
[[1]]   <env: 0x55a975b59410>
[[2]] $ <env: global>
```

```
env_parents(e2d)
```

```
[[1]]   <env: 0x55a978edcdd8>
[[2]] $ <env: empty>
```

可以利用 Use `parent.env()` 找到環境的 parent，但是 base 中沒有可以找出所有祖先的函數。

8.2.4 Getting and setting

存取環境中元素的方法和 list 類似: 使用 `$` 和 `[[`:

```
e3 <- env(x = 1, y = 2)
e3$x
```

```
[1] 1
```

```
e3$z <- 3
e3[["z"]]
```

```
[1] 3
```

但是不能使用 `[[` + 數字索引, 也不能單獨使用 `[`:

```
e3[[1]]
```

```
Error in e3[[1]]:
  取子集環境時的引數不正確
```

```
e3[c("x", "y")]
```

```
Error in e3[c("x", "y")]:
  'environment' 類型的物件無法具有子集合
```

當環境中的繫結不存在時 (簡單點, 就是變數不存在時)`$` 和 `[[` 會傳回 `NULL` 但不會引發錯誤, 如果要有錯誤警告, 則利用 `env_get()`：

```
e3$xyz
```

```
NULL
```

```
env_get(e3, "xyz")
```

```
Error in env_get(e3, "xyz"):
  找不到物件 'xyz'
```

當繫結不存在, 但是想要有預設值傳回時, 可以利用參數 `default` .

```
env_get(e3, "xyz", default = NA)
```

```
[1] NA
```

另有兩種方式可以在環境物件加入繫結:

- `env_poke()`[1] takes a name (as string) and a value:

  ```
  env_poke(e3, "a", 100)
  e3$a
  ```

  ```
  [1] 100
  ```

- `env_bind()` allows you to bind multiple values:

  ```
  env_bind(e3, a = 10, b = 20)
  env_names(e3)
  ```

  ```
  [1] "x" "y" "z" "a" "b"
  ```

`env_has()`: 是否環境中有繫結

```
env_has(e3, "a")
```

```
   a
TRUE
```

不能像是 list 中刪除元素的方式 (指派 `NULL` 給元素)，而必須使用 `env_unbind()`:

[1] You might wonder why rlang has `env_poke()` instead of `env_set()`. This is for consistency: `_set()` functions return a modified copy; `_poke()` functions modify in place.

```
e3$a <- NULL
env_has(e3, "a")
```

```
   a
TRUE
```

```
env_unbind(e3, "a")
env_has(e3, "a")
```

```
    a
FALSE
```

從一個物件 Unbinding 解除名稱，並不會刪除物件，是否刪除物件是 garbage collector 的工作.。可以參考GC.

See `get()`, `assign()`, `exists()`, and `rm()`. These are designed interactively for use with the current environment, so working with other environments is a little clunky. Also beware the `inherits` argument: it defaults to `TRUE` meaning that the base equivalents will inspect the supplied environment and all its ancestors.

8.2.5 Finalisers

Add something once rlang has an API. Also mention in data structures below

8.2.6 Advanced bindings

There are two more exotic variants of `env_bind()`:

- `env_bind_exprs()` creates **delayed bindings**, which are evaluated the first time they are accessed. Behind the scenes, delayed bindings create promises, so behave in the same way as function arguments.

  ```
  env_bind_exprs(current_env(), b = {Sys.sleep(1); 1})

  system.time(print(b))
  ```

  ```
  [1] 1
  ```

  ```
     user  system elapsed
        0       0       1
  ```

```
system.time(print(b))
```

```
[1] 1
```

```
   user  system elapsed 
  0.000   0.000   0.001 
```

Delayed bindings are used to implement `autoload()`, which makes R behave as if the package data is in memory, even though it's only loaded from disk when you ask for it.

- `env_bind_fns()` creates **active bindings** which are re-computed every time they're accessed:

```
env_bind_fns(current_env(), z1 = function(val) runif(1))

z1
```

```
[1] 0.0808
```

```
z1
```

```
[1] 0.834
```

The argument to the function allows you to also override behaviour when the variable is set:

```
env_bind_fns(current_env(), z2 = function(val) {
  if (missing(val)) {
    2
  } else {
     stop("Don't touch z2!", call. = FALSE)
  }
})

z2
```

```
[1] 2
```

```
z2 <- 3
```

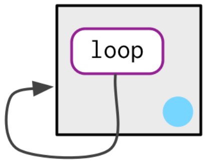
```
Error: Don't touch z2!
```

See `?delayedAssign()` and `?makeActiveBinding()`.

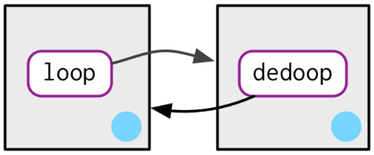
8.2.7 Exercises

1. List three ways in which an environment differs from a list.

2. Create an environment as illustrated by this picture.



3. Create a pair of environments as illustrated by this picture.



4. Explain why `e[[1]]` and `e[c("a", "b")]` don't make sense when `e` is an environment.

5. Create a version of `env_poke()` that will only bind new names, never re-bind old names. Some programming languages only do this, and are known as single assignment languages.

8.3 Recursing over environments

If you want to operate on every ancestor of an environment, it's often convenient to write a recursive function. This section shows you how, applying your new knowledge of environments to write a function that given a name, finds the environment `where()` that name is defined, using R's regular scoping rules.

The definition of `where()` is straightforward. It has two arguments: the name to look for (as a string), and the environment in which to start the search. (We'll learn why `caller_env()` is a good default in calling environments.)

```
where <- function(name, env = caller_env()) {
  if (identical(env, empty_env())) {
    # Base case
    stop("Can't find ", name, call. = FALSE)
  } else if (env_has(env, name)) {
    # Success case
    env
  } else {
    # Recursive case
    where(name, env_parent(env))
  }
}
```

3 個情況:

- The base case: 到達 empty environment 沒有 parent 無法繼續, 所以丟出 error.
- The successful case: 在 env 中找到 name ， 成功, 所以傳回 env。.
- The recursive case: 在 env 中找不到, 繼續在 parent 中找。.

These three cases are illustrated with these three examples:

```
where("yyy")
```

```
Error: Can't find yyy
```

```
x <- 5
where("x")
```

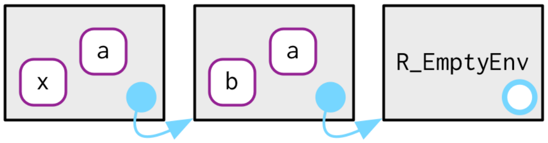
```
<environment: R_GlobalEnv>
```

```
where("mean")
```

```
<environment: base>
```

想像有兩個環境物件 (如圖):

```
e4a <- env(empty_env(), a = 1, b = 2)
e4b <- env(e4a, x = 10, a = 11)
```



- `where(a, e4a)` will find `a` in `e4a`.
- `where("b", e4a)` doesn't find `b` in `e4a`, so it looks in its parent, `e4b`, and finds it there.
- `where("c", e4a)` looks in `e4a`, then `e4b`, then hits the empty environment and throws an error.

It's natural to work with environments recursively, so `where()` provides a useful template. Removing the specifics of `where()` shows the structure more clearly:

```
f <- function(..., env = caller_env()) {
  if (identical(env, empty_env())) {
    # base case
  } else if (success) {
    # success case
  } else {
    # recursive case
    f(..., env = env_parent(env))
  }
}
```

Iteration vs recursion

也可以用迭代的方式

```
f2 <- function(..., env = caller_env()) {
  while (!identical(env, empty_env())) {
    if (success) {
      # success case
      return()
    }
    # inspect parent
    env <- env_parent(env)
  }

  # base case
}
```

8.3.1 Exercises

1. Modify `where()` to return *all* environments that contain a binding for `name`. Carefully think through what type of object the function will need to return.

2. Write a function called `fget()` that finds only function objects. It should have two arguments, `name` and `env`, and should obey the regular scoping rules for functions: if there's an object with a matching name that's not a function, look in the parent. For an added challenge, also add an `inherits` argument which controls whether the function recurses up the parents or only looks in one environment.

8.4 Special environments

這裡討論 package environments. 然後探討當函數建立時, 綁入函數的函數環境。還有當函數被呼叫時的執行環境 (ephemeral)。

套裝環境主要是看這些環境如何支援 namespaces。同時,namespace 讓 package 每次載入的時候, 都有一樣的行為, 而不售其他 packages 載入先後的影響。

8.4.1 Package environments and the search path

每個套件經由 `library()` 或 `require()` 接入成為總體環境的 parent。而最後一個接入的套件, 則是總體環境的第一個 parent:

load 和 attach 不一樣，當我們使用 library 的時候, 我們做的是 [^attach] 在環境串列中加入我們利用 library 載入的物件.. [^attach]: Note the difference between attached and loaded. A package is loaded automatically if you access one of its functions using `::`; it is only **attached** to the search path by `library()` or `require()`.

```
env_parent(global_env())
```

```
<environment: package:forcats>
attr(,"name")
[1] "package:forcats"
attr(,"path")
[1] "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5/forcats"
```

And the parent of that package is the second to last package you attached:

```
env_parent(env_parent(global_env()))
```

```
<environment: package:stringr>
attr(,"name")
[1] "package:stringr"
attr(,"path")
[1] "/home/linchao/R/x86_64-pc-linux-gnu-library/3.5/stringr"
```

如果一層一層 parent 回朔, 就可以到每個套件被接入的順序, 這也是 R 執行中會用到的 **search path** 因為這些環境的所有物件都可以經由 top-level interactive workspace 找到。

```
search_envs()
```

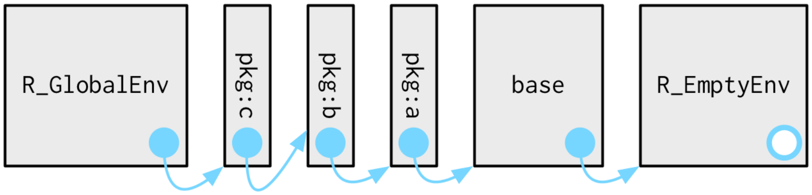
```
 [[1]] $ <env: global>
 [[2]] $ <env: package:forcats>
 [[3]] $ <env: package:stringr>
 [[4]] $ <env: package:purrr>
 [[5]] $ <env: package:readr>
 [[6]] $ <env: package:tibble>
 [[7]] $ <env: package:tidyverse>
 [[8]] $ <env: package:foreign>
 [[9]] $ <env: package:ggplot2>
[[10]] $ <env: package:babynames>
[[11]] $ <env: package:bindrcpp>
[[12]] $ <env: package:tidyr>
[[13]] $ <env: package:rlang>
[[14]] $ <env: package:usethis>
[[15]] $ <env: package:devtools>
[[16]] $ <env: package:dplyr>
[[17]] $ <env: package:moments>
[[18]] $ <env: mtcars>
[[19]] $ <env: mtcars>
[[20]] $ <env: package:stats>
... and 7 more environments
```

函數 `search()` 可以找出環境物件的名稱。

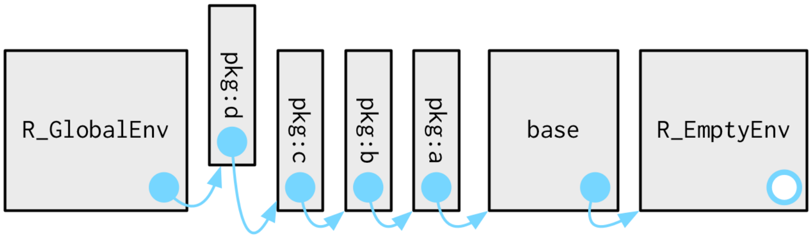
最後兩個環境物件都一樣:

- `Autoloads` 環境物件, 利用 delayed bindings 來節省記憶體, 也就是在需要的時候才載入 (loading)package 中的物件 (例如大型資料集)。
- base environment, `package:base` 或簡稱 `base`, 是 base 套裝的環境物件。用來載入其他套裝 (bootstrap)。利用函數 `base_env()` 存取.

利用圖型表示:



當利用 `library()` attach 其他套件的時候, 總體環境的 parent 馬上改變:



8.4.2 The function environment

當函數被建立的時候, 現有的環境會被繫結。稱為 *function environment*, 主要用來支援 lexical scoping. 在電腦語言中, 當函數紀錄它們的運作環境時, 我們說這個函數屬於 *closures*。, 這也是為甚麼這個字眼經常在 R 語言中出現。.

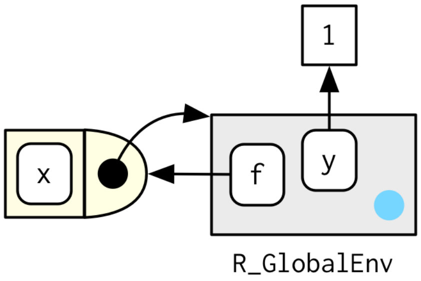
利用函數 `fn_env()` 可以得到函數的環境物件:

```r
y <- 1
f <- function(x) x + y
fn_env(f)
```

```
<environment: R_GlobalEnv>
```

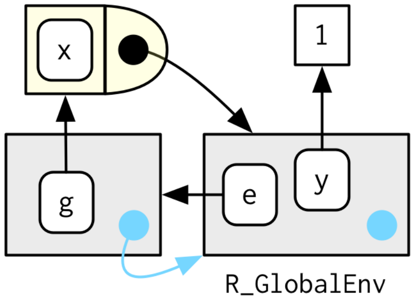
一樣利用函數 `environment(f)` 可以找到函數 `f` 的環境.

在圖形中, 函數被畫成類似子彈, 而彈頭的部分繫結環境。



在這個案例中,`f()` 繫結的環境物件, 就是繫結名稱 `f` 的環境。但並不一定總是這樣, 例如在下一個例子中,g 被繫結在新環境物件 `e` 中。但是函數 `g()` 繫結的是 global environment。這之間的分別是我們如何找到 `g` 和 `g` 如何找到他的變數。

```
e <- env()
e$g <- function() 1
```



8.4.3 Namespaces

在上面的圖形中, 我們已經知道套件的 parent 會隨著之前套件載入的順序不同而不同。這就導致 R 程式設計者必須保證個別套件上如果使用別的套件的函數，必須是原始目的的那一個。*namespaces* 就是為此目的而產生: 每個套件必須的使用必須一致, 而不管使用者如何載入套件.

以 `sd()` 為例子:

```
sd
```

```
function (x, na.rm = FALSE)
sqrt(var(if (is.vector(x) || is.factor(x)) x else as.double(x),
    na.rm = na.rm))
<bytecode: 0x55a971f03df8>
<environment: namespace:stats>
```

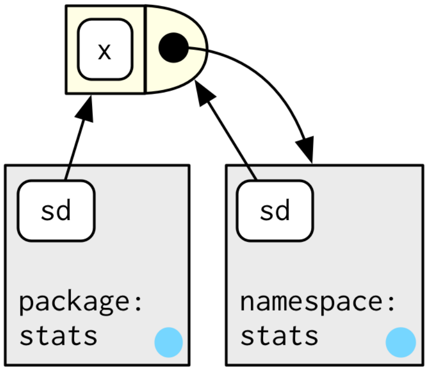
`sd()` 必須使用函數 `var()`, 因此這個 `var()` 到底來自 global environment, 還是其他接入 (attached) 的套件的這種問題必須避免。R avoids this problem by taking advantage of the function vs. binding environment described above.

每個套件中的函數和一對環境物件有關: 套件環境 (之前學到的) 還有 *namespace* 環境物件。

- package environment: 是套件的外部介面，這是 R 使用者如何在接入的套件中尋找函數的地方 (或者可以利用 `::`) package enviromnent 的 parent 由搜尋路徑決定 (可以利用 `search()` 知道) 決定 (也就是載入的順序)

- namespace environment: 是套件的內部介面。package environment 控制我們如何找到函數, 而 namespace environment 控制函數如何找到變數。

在 package environment 的每個繫結也可以在 namespace environment 中找到。這樣可以確保每個函數可以使用套件中的其他函數。但是有些繫結只能在 namespace 中找到 (例如內部或非輸出物件), 這種內部物件通常是用來隱藏一些繁瑣的且不需要給使用者看到的細節。;

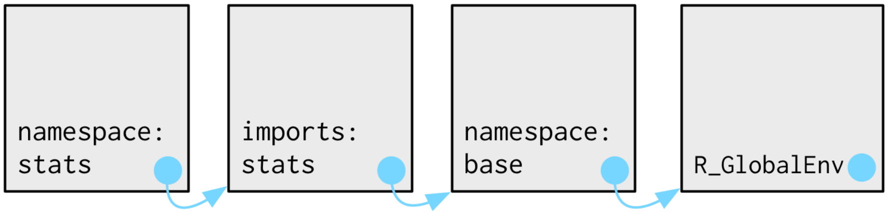


每個 namespace environment 有一樣的祖集合:

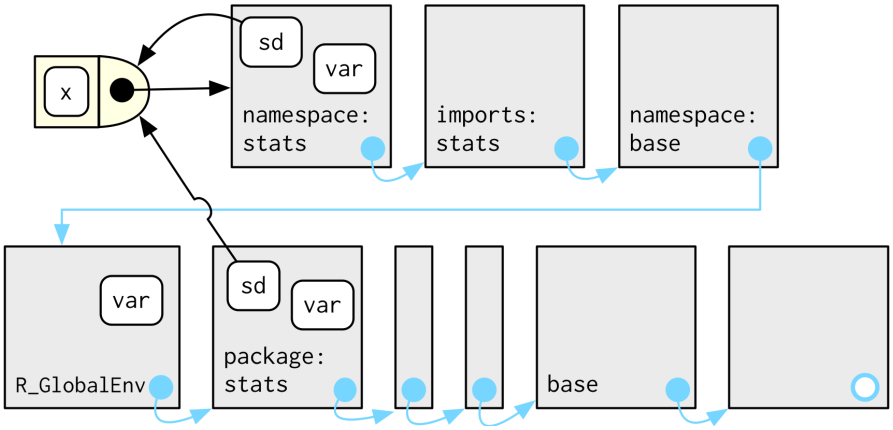
- 每個 namespace 有 *imports* 的環境物件，其中包含了套件中用到的所有函數繫結。而所謂輸入環境實際由套裝開發人員在檔案 `NAMESPACE` 指定
- 明確的輸入每個 `base` 函數, 很繁瑣, 所以 R 直接設定 `import enviroment` 的 parent 是 `base *namespae*`[^1]。

  The base namespace contains the same bindings as the base environment, but it has different parent.

- *base namespace* 的 parent 是總體環境 (#global environment). 參考下圖, 這種設計導致在 import environment 中找不到繫結時, 會開始再總體環境中尋找, 而之前提過總體環境通常是互動環境下名稱搜尋的開始路徑, 這也導致搜尋的方式受到套件載入順序的影響。因此,R 提供了了 `R CMD check` 來警告此種情況的發生。(雖然有麻煩, 但是由於 S3 方法的 dispatch 關係, 此種方式仍然留著)



綜合上述, 可以得到下圖::



所以當 sd() 搜尋 var 的值的時候, 搜尋順序是受到開發者的指定 (在檔案 NAMESPACE 利用 import), 而不會受到套裝使用者的影響。這樣保證每次套件程式碼執行的時候, 都一樣, 而不會受到一般使用者載入套件的順序而影響。

注意在 package 和 namespace 兩種環境之間沒有直接的連結. 連結是由函數環境定義。

8.4.4 Execution environments

execution environment. 下面的函數第一次執行的時候會傳回甚麼? 第 2 次呢?

```
g <- function(x) {
  if (!env_has(current_env(), "a")) {
    message("Defining a")
    a <- 1
  } else {
    a <- a + 1
  }
  a
}
```

再一次利用下面的調用, 確認你的答案:

```
g(10)
```

```
Defining a
```

```
[1] 1
```

```
g(10)
```

```
Defining a
```

```
[1] 1
```

```
g(11)
```

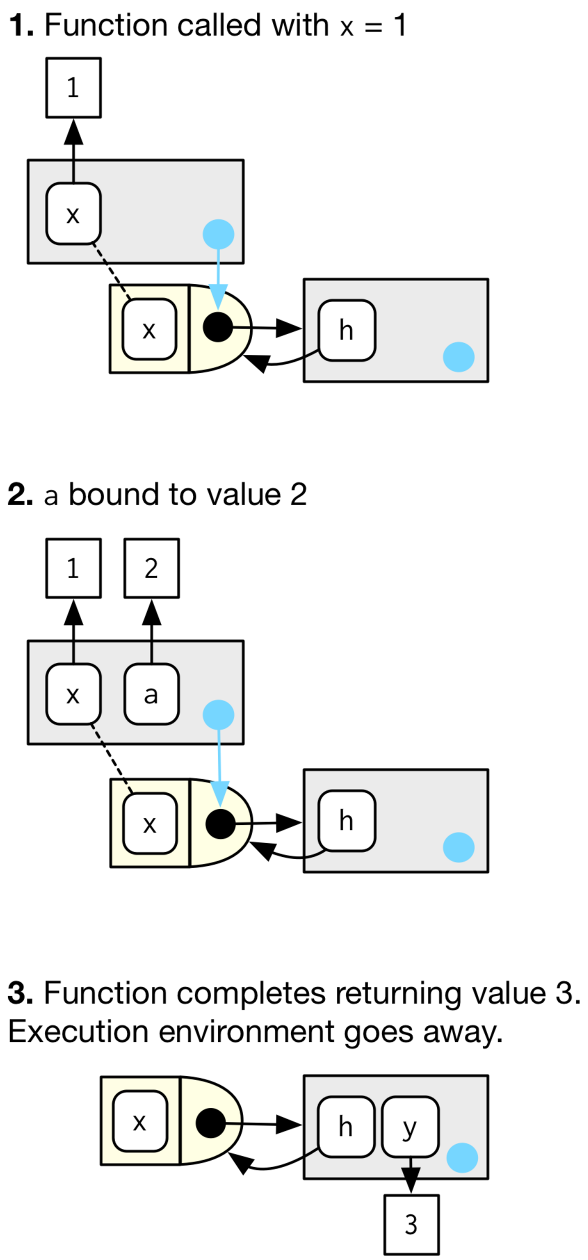
```
Defining a
```

```
[1] 1
```

這個函數每次執行都傳回一樣的答案, 參考a fresh start. 每次函數被調用的時候, 一個新的環境都會被建立來主導執行。這種環境稱為執行環境。而執行環境的 parent 為 function environment.

用另一個簡單點的例子說明. (圖中, 執行環境的 parent 間接表示: 經由函數環境).

```
h <- function(x) {
  # 1.
  a <- 2 # 2.
  x + a
}
y <- h(1) # 3.
```



1. Function called with x = 1

2. a bound to value 2

3. Function completes returning value 3.
Execution environment goes away.

執行環境 (execution environment) 短暫存在, 當函數執行完畢通常會被 GC。在幾種情況下, 會在記憶體存在比較久, 第一種是回傳給另一個變數:

```r
h2 <- function(x) {
  a <- x * 2
  current_env()
}

e <- h2(x = 10)
env_print(e)
```

```
<environment: 0x55a9752403a8>
parent: <environment: global>
bindings:
 * a: <dbl>
 * x: <dbl>
```

```r
fn_env(h2)
```

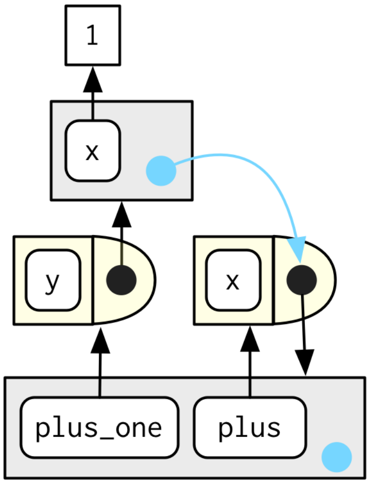
```
<environment: R_GlobalEnv>
```

Another way to capture it is to return an object with a binding to that environment, like a function. The following example illustrates that idea with a function factory, `plus()`. We use that factory to create a function called `plus_one()`.

There's a lot going on in the diagram because the enclosing environment of `plus_one()` is the execution environment of `plus()`.

```r
plus <- function(x) {
  function(y) x + y
}

plus_one <- plus(1)
plus_one
```

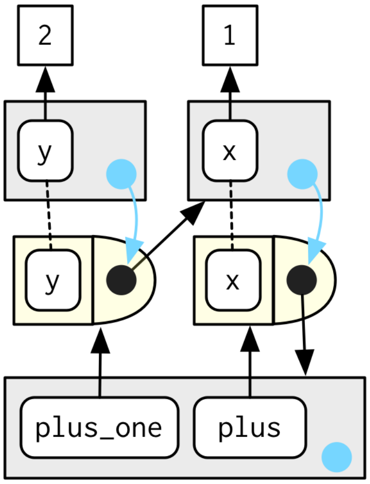
```
function(y) x + y
<environment: 0x55a9735c70c8>
```



What happens when we call `plus_one()`? Its execution environment will have the captured execution environment of `plus()` as its parent:

```
plus_one(2)
```

```
[1] 3
```



You'll learn more about function factories in functional programming.

8.4.5 Exercises

1. How is `search_envs()` different to `env_parents(global_env())`?

2. Draw a diagram that shows the enclosing environments of this function:

```
f1 <- function(x1) {
  f2 <- function(x2) {
    f3 <- function(x3) {
      x1 + x2 + x3
    }
    f3(3)
  }
  f2(2)
}
f1(1)
```

3. Write an enhanced version of `str()` that provides more information about functions. Show where the function was found and what environment it was defined in.

8.5 The call stack

還有另一種環境稱為 **caller** environment, 可以經由 `rlang::caller_env()` 存取。. This provides the environment from which the function was called, and hence varies based on how the function is called, not how the function was created. As we saw above this is a useful default whenever you write a function that takes an environment as an argument.

`parent.frame()` is equivalent to `caller_env()`; just note that it returns an environment, not a frame.

To fully understand the caller environment we need to discuss two related concepts: the **call stack**, which is made up of **frames**. Executing a function creates two types of context. You've learned about one already: the execution environment is a child of the function environment, which is determined by where the function was created. There's another type of context created by where the function was called: this is called the call stack.

There are also a couple of small wrinkles when it comes to custom evaluation. See environments vs. frames for more details.

8.5.1 Simple call stacks

Let's illustrate this with a simple sequence of calls: `f()` calls `g()` calls `h()`.

```
f <- function(x) {
  g(x = 2)
}
g <- function(x) {
  h(x = 3)
}
h <- function(x) {
  stop()
}
```

The way you most commonly see a call stack in R is by looking at the `traceback()` after an error has occured:

```
f(x = 1)
#> Error:
traceback()
#> 4: stop()
#> 3: h(x = 3)
#> 2: g(x = 2)
#> 1: f(x = 1)
```

Instead of `stop()` + `traceback()` to understand the call stack, we're going to use `lobstr::cst()` to print out the **c**all **s**tack **t**ree:

```
h <- function(x) {
  lobstr::cst()
}
f(x = 1)
#> ???
#> ?? ? f(x = 1)
#>   ?? ? g(x = 2)
#>     ?? ? h(x = 3)
#>       ?? ? lobstr::cst()
```

This shows us that `cst()` was called from `h()`, which was called from `g()`, which was called from `f()`. Note that the order is the opposite from `traceback()`. As the call stacks get more compliated, I think it's easier to understand the sequence of calls if you start from the beginning, rather than the end (i.e. `f()` calls `g()`; rather than `g()` was called by `f()`).

8.5.2 Lazy evaluation

The call stack above is simple - while you get a hint that there's some tree-like structure involved, everything happens on a single branch. This is typical of a call stack when all arguments are eagerly evaluated.

Let's create a more complicated example that involves some lazy evaluation. We'll create a sequence of functions, `a()`, `b()`, `c()`, that pass along an argument `x`.

```
a <- function(x) b(x)
b <- function(x) c(x)
c <- function(x) x

a(f())
#> ???
#> ?? ? a(f())
#> ??? ?? ? b(x)
#> ???   ?? ? c(x)
#> ?? ? f()
#>   ?? ? g(x = 2)
#>     ?? ? h(x = 3)
#>       ?? ? lobstr::cst()
```

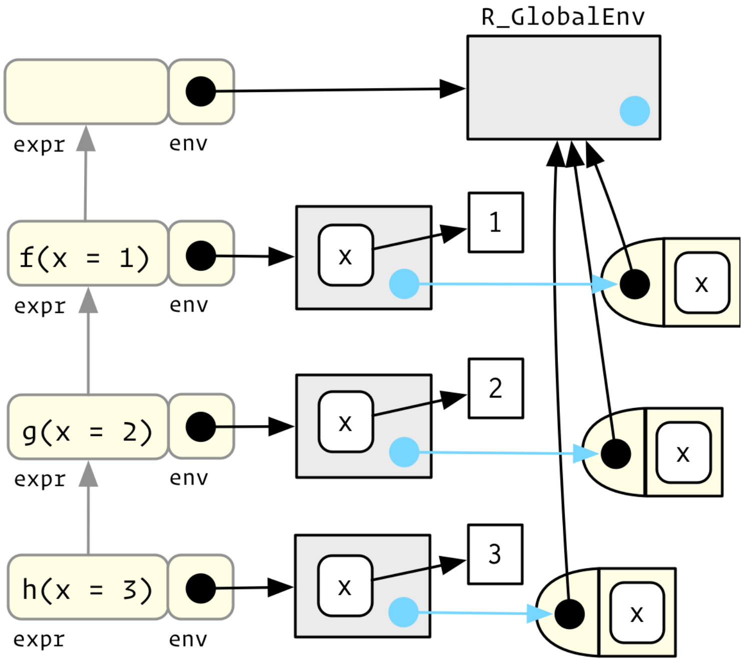
`x` is lazily evaluated so this tree gets two branches. In the first branch `a()` calls `b()`, then `b()` calls `c()`. The second branch starts when `c()` evaluates its argument `x`. This argument is evaluated in a new branch because the environment in which it is evaluated is the global environment, not the environment of `c()`.

8.5.3 Frames

Each element of the call stack is a **frame**[2], also known as an evaluation context. The frame is an extremely important internal data structure, and R code can only access a small part of the data structure because it's so critical. A frame has three main components that are accessible from R:

- An expression (labelled with `expr`) giving the function call. This is what `traceback()` prints out.
- An environment (labelled with `env`), which is typically the execution environment of a function. There are two main exceptions: the environment of the global frame is the global environment, and calling `eval()` also generates frames, where the environment can be anything.
- A parent, the previous call in the call stack (shown by a grey arrow).

[2] NB: `?environment` uses frame in a different sense: "Environments consist of a *frame*, or collection of named objects, and a pointer to an enclosing environment.". We avoid this sense of frame, which comes from S, because it's very specific and not widely used in base R. For example, the "frame" in `parent.frame()` is an execution context, not a collection of named objects.



(To focus on the calling environments, I have omitted the bindings in the global environment from `f`, `g`, and `h` to the respective function objects.)

The frame also holds exit handlers created with `on.exit()`, restarts and handlers for the condition system, and which context to `return()` to when a function completes. These are important for the internal operation of R, but are not directly accessible.

8.5.4 Dynamic scope

Looking up variables in the calling stack rather than in the enclosing environment is called **dynamic scoping**. Few languages implement dynamic scoping (Emacs Lisp is a notable exception.) This is because dynamic scoping makes it much harder to reason about how a function operates: not only do you need to know how it was defined, you also need to know the context in which it was called. Dynamic scoping is primarily useful for developing functions that aid interactive data analysis. It is one of the topics discussed in non-standard evaluation.

8.5.5 Exercises

1. Write a function that lists all the variables defined in the environment in which it was called. It should return the same results as `ls()`.

8.6 As data structures

As well as powering scoping, environments are also useful data structures in their own right because they have reference semantics. There are three common problems that they can help solve:

- **Avoiding copies of large data**. Since environments have reference semantics, you'll never accidentally create a copy. This makes it a useful vessel for large objects. Bare environments are not that pleasant to work with; I recommend using R6 objects instead. Learn more in [R6].

- **Managing state within a package**. Explicit environments are useful in packages because they allow you to maintain state across function calls. Normally, objects in a package are locked, so you can't modify them directly. Instead, you can do something like this:

  ```
  my_env <- new.env(parent = emptyenv())
  my_env$a <- 1

  get_a <- function() {
    my_env$a
  }
  set_a <- function(value) {
    old <- my_env$a
    my_env$a <- value
    invisible(old)
  }
  ```

  Returning the old value from setter functions is a good pattern because it makes it easier to reset the previous value in conjunction with `on.exit()` (see more in on exit).

- **As a hashmap**. A hashmap is a data structure that takes constant, O(1), time to find an object based on its name. Environments provide this behaviour by default, so can be used to simulate a hashmap. See the CRAN package hash for a complete development of this idea.

8.7 <<-

The ancestors of an environment have an important relationship to `<<-`. The regular assignment arrow, `<-`, always creates a variable in the current environment. The deep assignment arrow, `<<-`, never creates a variable in the current environment, but instead modifies an existing variable found by walking up the parent environments.

```r
x <- 0
f <- function() {
  x <<- 1
}
f()
x
```

```
[1] 1
```

If `<<-` doesn't find an existing variable, it will create one in the global environment. This is usually undesirable, because global variables introduce non-obvious dependencies between functions. `<<-` is most often used in conjunction with a closure, as described in Closures.

8.7.1 Exercises

1. What does this function do? How does it differ from `<<-` and why might you prefer it?

```r
rebind <- function(name, value, env = caller_env()) {
  if (identical(env, empty_env())) {
    stop("Can't find `", name, "`", call. = FALSE)
  } else if (env_has(env, name)) {
    env_poke(env, name, value)
  } else {
    rebind(name, value, env_parent(env))
  }
}
rebind("a", 10)
```

```
Error: Can't find `a`
```

```r
a <- 5
rebind("a", 10)
a
```

```
[1] 10
```

8.8 Quiz answers

1. There are four ways: every object in an environment must have a name; order doesn't matter; environments have parents; environments have reference semantics.
2. The parent of the global environment is the last package that you loaded. The only environment that doesn't have a parent is the empty environment.
3. The enclosing environment of a function is the environment where it was created. It determines where a function looks for variables.
4. Use `caller_env()` or `parent.frame()`.
5. `<-` always creates a binding in the current environment; `<<-` rebinds an existing name in a parent of the current environment.

8.8.1 term

8.8.1.0.1 global environment : 總體環境

8.8.1.0.2 package environments

8.8.1.0.3 imports environment

Chapter 9

基本繪圖

9.1 reference

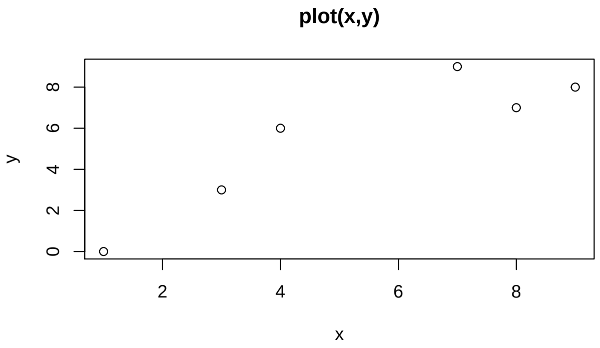
Quric R

一般視窗作業系統的繪圖，會有一個抽象物件 device 在這個 device 上，有一個抽象物件 canvas，和繪圖工具例如，水彩筆之類的抽象物件可以指定顏色，線條粗細等等。

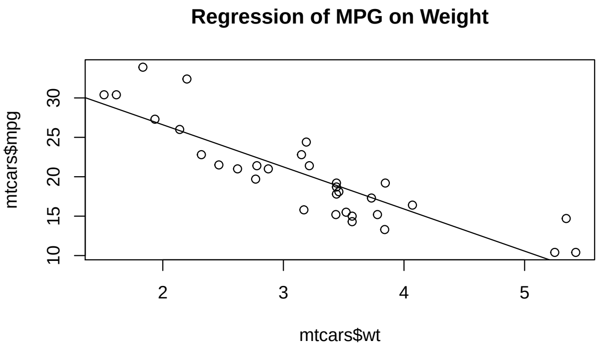
9.2 basic

基本繪圖函數，plot(),hist()
### plot

```
x <- c(1,3,4,7,8,9)
y <- c(0,3,6,9,7,8)
plot(x,y,main='plot(x,y)')
```



```
plot(mtcars$wt, mtcars$mpg)
abline(lm(mtcars$mpg~mtcars$wt))
title("Regression of MPG on Weight")
```



```
attach(mtcars)
plot(wt, mpg)
abline(lm(mpg~wt))
title("Regression of MPG on Weight")
```

plot() 函數，新建視窗然後畫出 weight vs. miles per gallon.
abline() 不會開啓新視窗做圖，abline() 的語法如下:

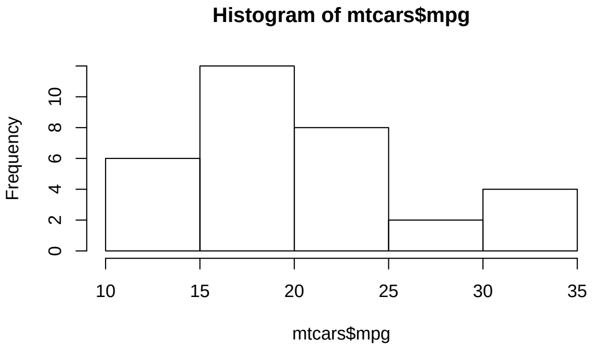
abline(a = NULL, b = NULL, h = NULL, v = NULL, reg = NULL,coef = NULL, untf = FALSE, ...)
代表的意思是畫出一條直線 a+bx ，其中 a 代表截距常數，b 則是斜率常數。例如 Y=2X+1，則對應的函數為 abline(a=1,b=2) 。

9.2.1 Histograms

函數 hist(x)：其中
x 是一個數字向量，選項 freq=FALSE 則是畫出 probability densities 而不是次數 frequencies.
選項：breaks= ## 則是控制分成幾份。

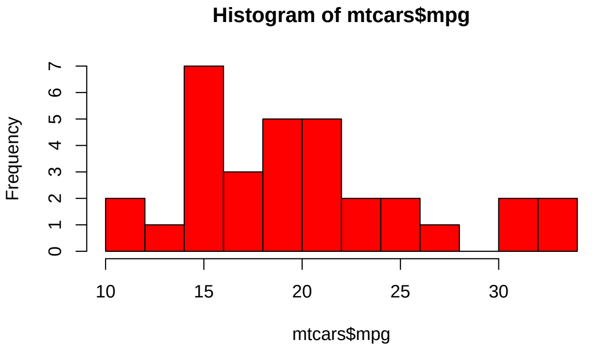
Simple Histogram

```
hist(mtcars$mpg)
```



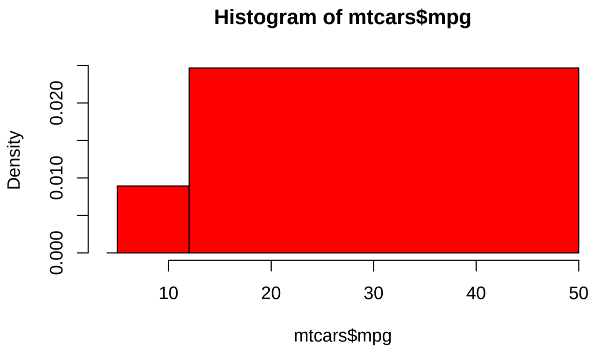
分成 **12** 份，顏色紅色。

```r
hist(mtcars$mpg, breaks=12, col="red")
```



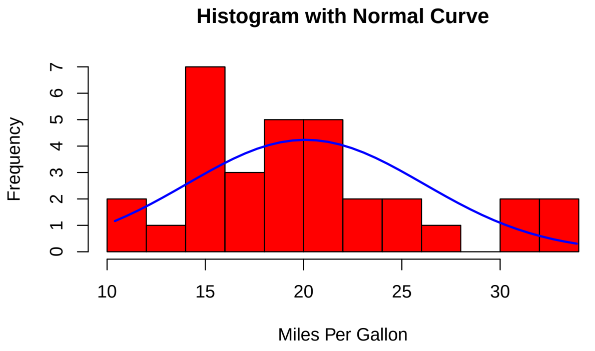
自行給定分割點

```r
hist(mtcars$mpg, breaks=c(4,5,12,50), col="red")
```



9.2.1.1 Add a Normal Curve

```
x <- mtcars$mpg
h<-hist(x, breaks=10, col="red", xlab="Miles Per Gallon",
    main="Histogram with Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



hint: 因為 $pdf = \frac{\frac{n}{N}}{bin \quad width}$ 所以 $n = pdf \times N \times (bin \quad width)$

Histograms can be a poor method for determining the shape of a distribution because it is so strongly affected by the number of bins used.

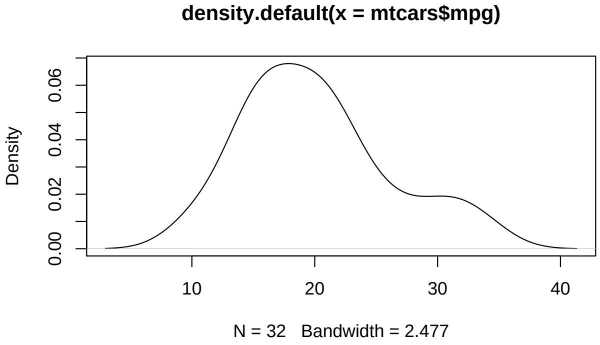
```
h
```

```
$breaks
 [1] 10 12 14 16 18 20 22 24 26 28 30 32 34

$counts
 [1] 2 1 7 3 5 5 2 2 1 0 2 2

$density
 [1] 0.0312 0.0156 0.1094 0.0469 0.0781 0.0781 0.0312 0.0312 0.0156
[10] 0.0000 0.0312 0.0312

$mids
 [1] 11 13 15 17 19 21 23 25 27 29 31 33

$xname
[1] "x"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

9.2.2 Kernel Density Plot

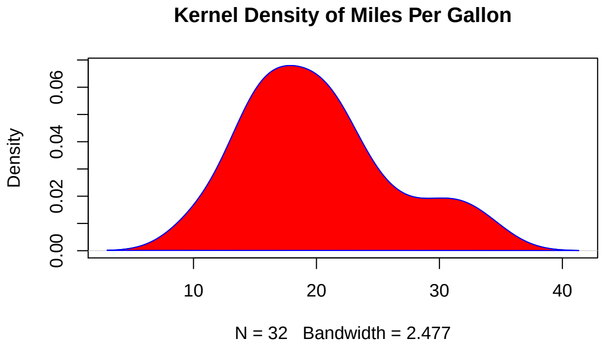
核密度畫圖 (Kernal density plots) 用來觀察一個變量，比較有有效率，指令如: `plot(density(x))` 其中 x 是一個數字向量。

```
d <- density(mtcars$mpg) # returns the density data
plot(d) # plots the results
```



9.2.2.1 Filled Density Plot

```
d <- density(mtcars$mpg)
plot(d, main="Kernel Density of Miles Per Gallon")
polygon(d, col="red", border="blue")
```



9.2.3 Saving Graphs

可以利用 menu 存檔, 例如 File -> Save As. 也可以利用函數:

Function	Output to
pdf(“mygraph.pdf”)	pdf file
win.metafile(“mygraph.wmf”)	windows metafile
png(“mygraph.png”)	png file
jpeg(“mygraph.jpg”)	jpeg file
bmp(“mygraph.bmp”)	bmp file
postscript(“mygraph.ps”)	postscript file

9.3 多個圖和疊圖

傻瓜指令例如 (plot, hist, boxplot, etc.) 基本上會開啓新視窗, 然後畫圖。這裡討論幾種自行控制的方法:
多視窗 (multiple windows)、覆圖 (combining figure) 和疊圖 (overlay) 討論。

9.3.1 多視窗

開多視窗的方法根據 OS 而不同，如下：

Function	Platform
windows()	Windows
X11()	Unix
quartz()	Mac

要關閉視窗，可以用函數 `dev.off()`
如果要知道目前的視窗是那一個，可以參考 `dev.cur()`。Note：如果目前有多個圖窗，則一直呼叫這個函數 `dev.off()`，會依次關閉視窗，直到函數傳回 NULL。

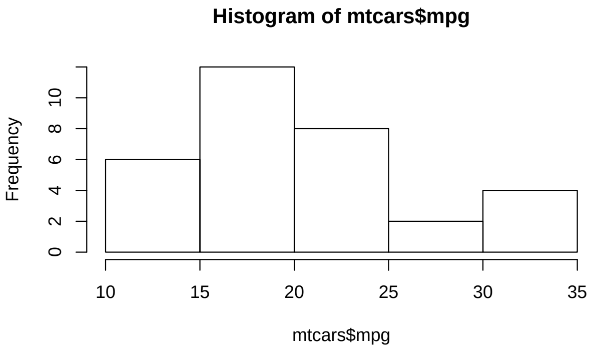
驗證上面的說法可以利用

```
hist(mtcars$mpg)
dev.cur()
```

```
pdf
  2
```

```
hist(mtcars$mpg)
dev.cur()
```

```
pdf
  2
```



上面的範例，可以看到圖裝置的編號都是 2，而且最後只有一個圖。

```
hist(mtcars$mpg)
dev.cur()
```

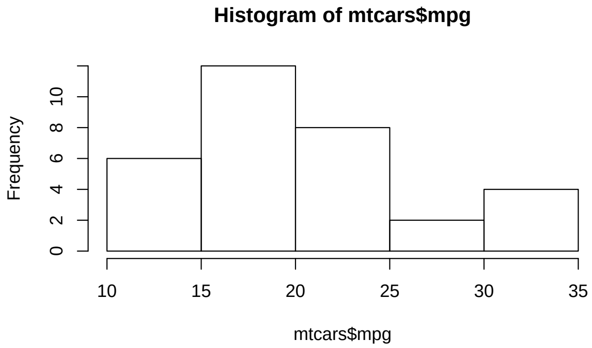
```
pdf
  2
```

```
hist(mtcars$mpg)
dev.cur()
```

```
pdf
  2
```

```
dev.off()
```

```
null device
          1
```



討論下面的輸出結果，我只看到一個視窗：hint：無法在 rmarkdown 執行？確認，在 rscript 中執行正常。

```
X11()
hist(mtcars$mpg)
dev.cur()
X11()
hist(mtcars$mpg)
dev.cur()
```

測試

```
plot(1:1)
dev.new()
plot(2,2)
dev.set(dev.prev()) # go back to first
title(main="test dev 1")

dev.set(dev.next()) # go to second
title(main="test dev 2")
```

9.3.2 合併 Plots

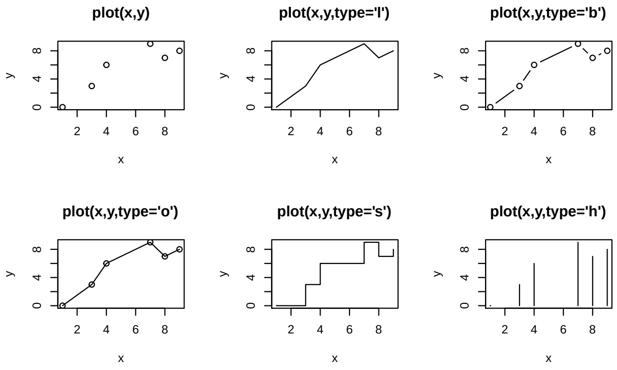
同時顯示多個 plot 的結果，可以利用函數 par() or layout( )。

函數 par( ) 的選項有：mfrow=c(nrows, ncols) 填 plot 的方向為橫行。mfcol=c(nrows, ncols) 填入的方向為直行。

9.3.2.1 par()

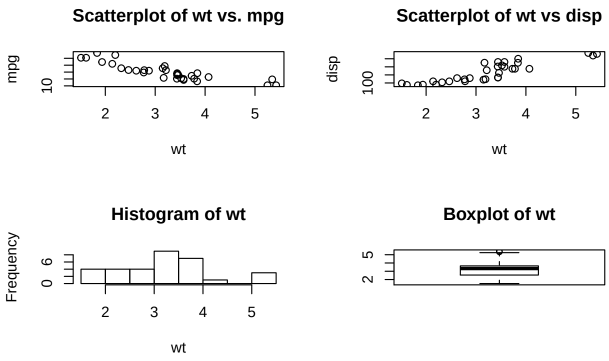
about par()

範例



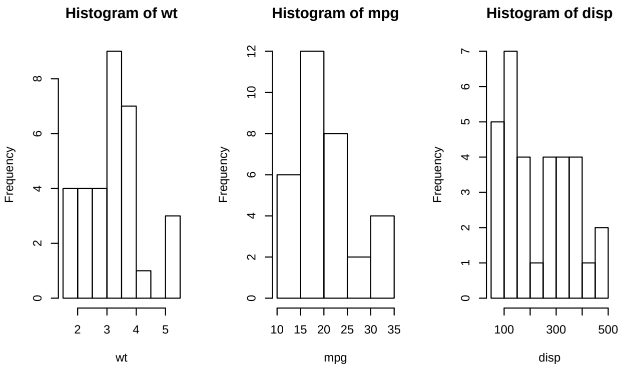
範例

```
#attach(mtcars)

par(mfrow=c(2,2))
plot(wt,mpg, main="Scatterplot of wt vs. mpg")
plot(wt,disp, main="Scatterplot of wt vs disp")
hist(wt, main="Histogram of wt")
boxplot(wt, main="Boxplot of wt")
```



1x3 layout

```
# attach(mtcars)
par(mfrow=c(1,3))
hist(wt)
hist(mpg)
hist(disp)
#dev.off()
```



note: par(mar) 列出 margin
par(mar=c(1,1,1,1)) 更動 margin

plot.new() 常發生的錯誤: figure margins too large

有兩個原因: 1 是畫布過小 2, 當前畫布的上下左右距離過大這裡看看如何解第二個發生原因

默認的畫布上邊款的距離為: 預設為 c(5, 4, 4, 2) + 0.1. 對應 c(bottom, left, top, right), 也就是順時針, 由下到右結束一圈。

我們可以將其設置為 0, 同時將目前的設定紀錄, 後面程式的第一行

```
op <- par(mar = rep(0, 4))    # op 會紀錄之前的 margin = 5.1 4.1 4.1 2.1
plot.new() # 畫圖
par(op) # 改回原先的 margin
```

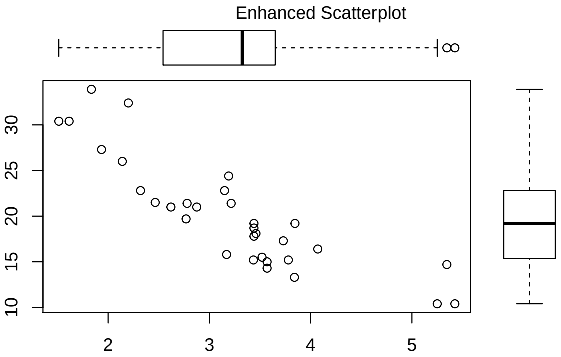
是不是有需要寫個新函數, 用來重設裝置參數? 下面這個函數 `resetPar()` 重新設定 device: 主要試看這一行:`par(no.readonly = TRUE)`。意思是說, 不要紀錄 readonly 的設定。

```
resetPar <- function() {
    dev.new() # 重新開始一個裝置
    op <- par(no.readonly = TRUE)# 因為是新的, 所以裡面的設定都是預設值
    dev.off()# 只是把之前的新裝置關掉
    op # 傳回預設值
}
```

其他用法 par() # view current settings opar <- par() # make a copy of current settings par(col.lab=“red”) # red x and y labels hist(mtcars$mpg) # create a plot with these new settings par(opar) # restore original settings

更多的圖形控制: 在散佈圖中加入 boxplot

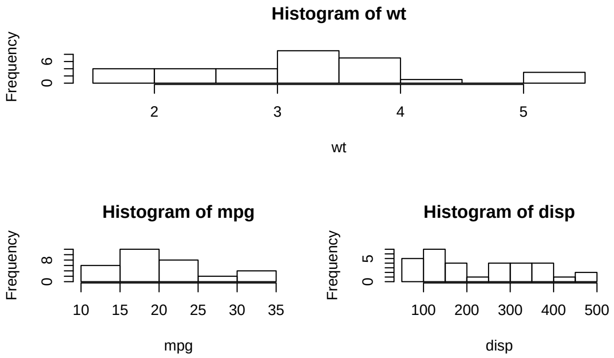
```
par(mar=rep(2,4))
# 整個圖的座標, 想成左下角 (0,0) 右上角 (1,1)
par(fig=c(0,0.8,0,0.8))# 左下 (0,0) 右上 (0.8,0.8)
# 左下 (x1,y1) 右上 (x2,y2) => c(x1,x2,y1,y2)
plot(mtcars$wt, mtcars$mpg, xlab="Car Weight",
  ylab="Miles Per Gallon")
par(fig=c(0,0.8,0.55,1), new=TRUE)
boxplot(mtcars$wt, horizontal=TRUE, axes=FALSE)
par(fig=c(0.65,1,0,0.8),new=TRUE)
boxplot(mtcars$mpg, axes=FALSE)
mtext("Enhanced Scatterplot", side=3, outer=TRUE, line=-3)
```



9.3.2.2 `layout()`

函數 layout( ) 的使用方法為 layout(mat) 其中 mat 的元素用來指定圖形號碼。例如分成 4 個格子, 順序為左右上下 (byrow=TRUE) 1 在第一 ROW, 占用 [1,1]-[1,2],2,3, 分別占用 [2,1] 和 [2,2]

```
# One figure in row 1 and two figures in row 2

#attach(mtcars)
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(wt)
hist(mpg)
hist(disp)
#dev.off()
```

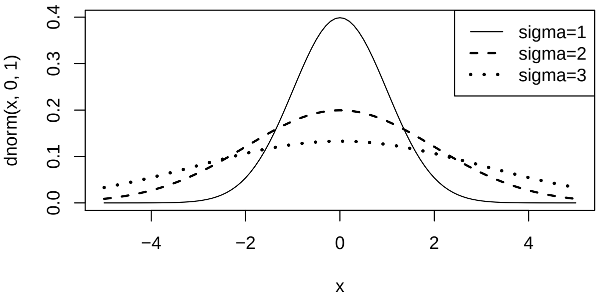


在 `layout()` 函數中，也可以更改圖形大小，其參數為：widths= 數字向量，用來代表 column 寬度 heights= 數字向量，用來代表 row 高度

note:Relative widths are specified with numeric values. Absolute widths (in centimetres) are specified with the lcm() function.

9.3.3 疊圖

```
curve( dnorm(x,0,1), -5 , 5, lwd=1, lty=1)
curve( dnorm(x,0,2),add=TRUE, lwd=2, lty=2)
curve( dnorm(x,0,3) , add=TRUE, lwd=3, lty=3)
# Add the legend
legend( "topright",c("sigma=1","sigma=2","sigma=3") , lwd=1:3, lty=1:3)
```



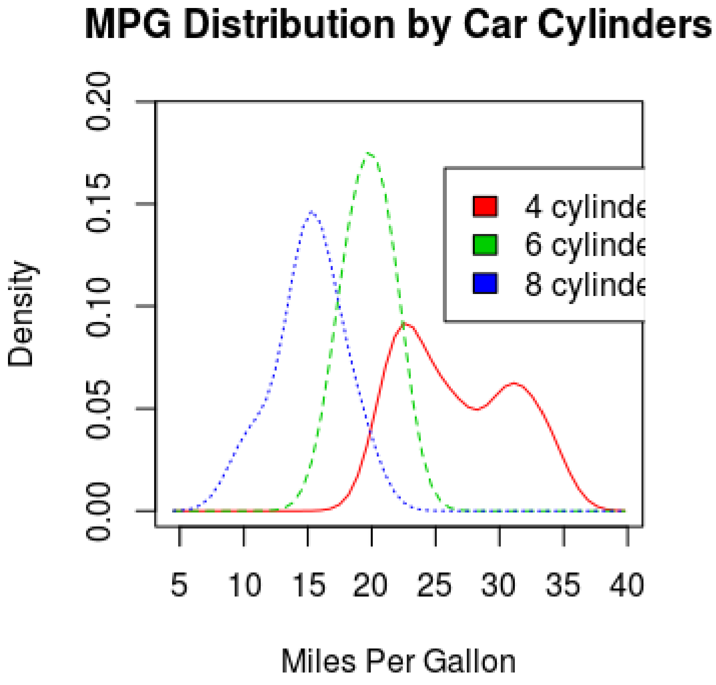
9.4 分組畫圖

9.4.1 Comparing Groups VIA Kernal Density

The sm.density.compare( ) function in the sm package allows you to superimpose the kernal density plots of two or more groups. The format is sm.density.compare(x, factor) where x is a numeric vector and factor is the grouping variable.

9.4.1.1 Compare MPG distributions for cars with 4,6, or 8 cylinders

```
library(sm)
attach(mtcars)

#create value labels
cyl.f <- factor(cyl, levels= c(4,6,8),
  labels = c("4 cylinder", "6 cylinder", "8 cylinder"))

#plot densities
sm.density.compare(mpg, cyl, xlab="Miles Per Gallon")
title(main="MPG Distribution by Car Cylinders")

#add legend via mouse click
colfill<-c(2:(2+length(levels(cyl.f))))
legend(locator(1), levels(cyl.f), fill=colfill)
```

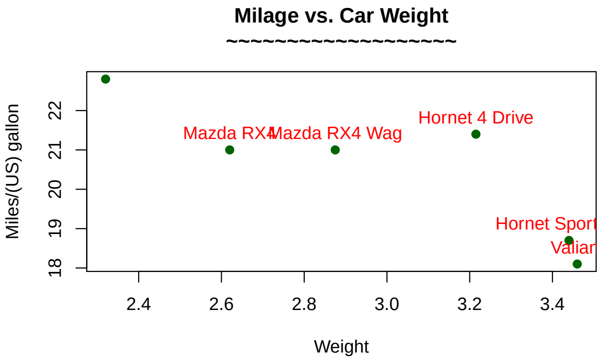


9.5 Add texts within the graph

The text() function can be used to draw text inside the plotting area. A simplified format of the function is :

text(x, y, labels) x and y: 文字座標; labels: 例如 “a label” pos: 下左上右，1234 cex: 放大倍數，例如，0.65。範例:

```
d<-head(mtcars)
plot(d[,'wt'], d[,'mpg'],
     main="Milage vs. Car Weight\n~~~~~~~~~~~~~~~~~~~",
      xlab="Weight", ylab="Miles/(US) gallon",
      pch=19, col="darkgreen")
text(d[,'wt'], d[,'mpg'],  row.names(d),     cex=1,pos=3,col="red")
```



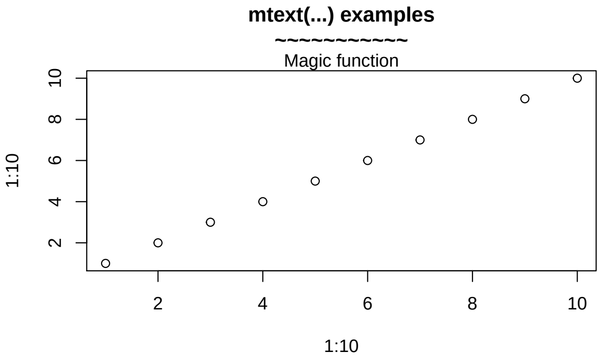
9.5.1 Add text in the margins of the graph

在圖形周圍給文字:

mtext(text, side=3) text：例如 “a label” side：哪一側:
順時針 1: 下 2: 左 3: 上 4: 又

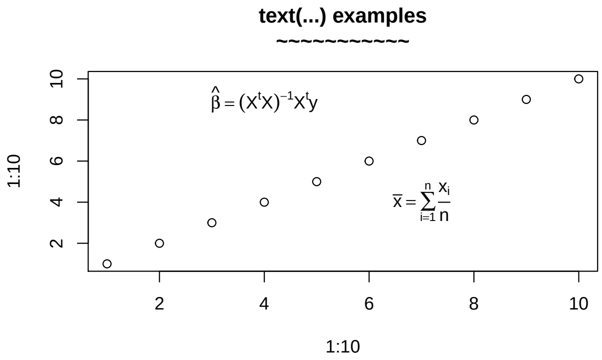
範例:

```
plot(1:10, 1:10,
     main="mtext(...) examples\n~~~~~~~~~~~")
mtext("Magic function", side=3)
```



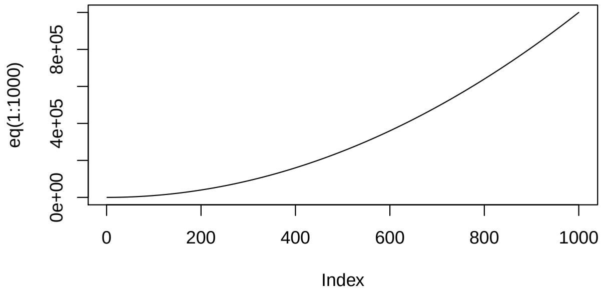
9.5.2 Add mathematical annotation to a plot

```
plot(1:10, 1:10,
     main="text(...) examples\n~~~~~~~~~~~~")
text(4, 9, expression(hat(beta) == (X^t * X)^{-1} * X^t * y))
text(7, 4, expression(bar(x) == sum(frac(x[i], n), i==1, n)))
```



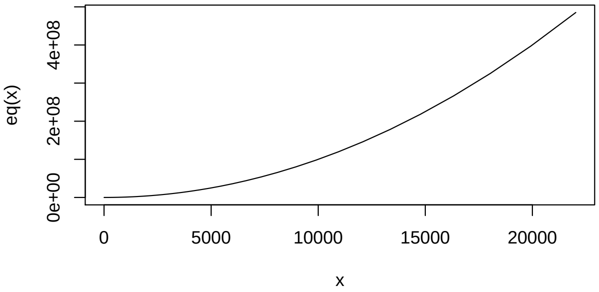
9.6 函數畫圖

```
eq = function(x){x*x}
plot(eq(1:1000), type='l')
```

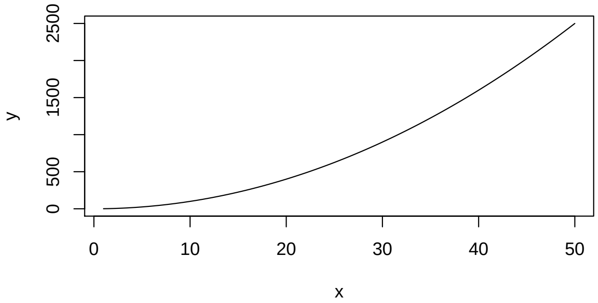


問題是如果 x 座標的增加不是 1 單位?

```
x<-seq(1,10,0.1)
y<-exp(x)
x<-y

eq = function(x){x*x}
plot(x,eq(x), type='l')
```



```
eq = function(x){x*x}
curve(eq, from=1, to=50, xlab="x", ylab="y")
```



問題: 解釋為何錯誤

```
eq = function(x){x*x}
y<-eq(1:50)
curve(y, xlab="x", ylab="y")
```

```
Error in y(x):
  沒有這個函數 "y"
```

問題: 如何修正下面的錯誤?

```r
eq = function(x){x*x}
z<-1:50

curve(eq(z), xlab="x", ylab="y")
```

solution:

Chapter 10

Sample and Distribution 01

## 隨機抽樣函數 sample(x,n,replace=FALSE). 其中 x 為要抽取的向量, n 為樣本容量. replace 預設為 false

1. no replacement, 等機率:

例如從 52 張撲克牌中抽取 5 張:

```
sample(1:52, 5)
```

```
[1]  5 43 31  8  1
```

2. replacement: 例如拋一枚均勻的硬幣 10 次

```
 sample(c("H", "T"), 10, replace=T)
```

```
 [1] "H" "H" "H" "T" "T" "T" "H" "H" "H" "H"
```

練習: 一棵骰子擲 10 次可表示為:

3) 不等可能的隨機抽樣: sample(x, n, replace=TRUE, prob=y) prob=y 指定 x 中元素出現的概率, 向量 y 與 x 等長度. 例如一娃娃機取出成功的概率為 0.6, 那麼 10 次的試驗為:

```
sample(c("sucess", "fail"), 10, replace=T, prob=c(0.6,0.4))
```

```
 [1] "fail"   "sucess" "sucess" "fail"   "fail"   "sucess" "sucess"
 [8] "sucess" "fail"   "sucess"
```

10.1 排列組合與概率的計算

例從一副 52 張撲克中取 4 張, 求以下事件的概率:
1. 抽取的 4 張依次為紅心 A, 方塊 A, 黑桃 A 和梅花 A 的概率;
2. 一次抽取 4 張為紅心 A, 方塊 A, 黑桃 A 和梅花 A 的概率.

his summation expression $\sum_{i=1}^{n} X_i$ appears inline.

解
1) 抽取的 4 張是有次序的, 因此使用排列來求解. 所求的事件 (記為 A) 概率為 $P(A) = \frac{1}{52\times51\times50\times49}$ 利用 R 函數

```
1/prod(52:49)
```

```
[1] 1.54e-07
```

2. 沒有次序的, 可以使用組合數來求解.

$$P(B) = \frac{1}{(52,4)}$$

其中 $(n,m) = \frac{n!}{m!(n-m)!}$, 可以利用函數 choose(), 例如

```
1/choose(52,4)
```

```
[1] 3.69e-06
```

10.2 distribution

標準表格上下沒有線條, 左右有

名稱	R 函數	選項
beta	beta	shape1, shape2
binomial	binom	size, prob
Cauchy	cauchy	location=0, scale=1
chi-sqaured (χ^2)	chisq	df, ncp
exponential	exp	rate
Fisher (F)	f	df1, df2, ncp
gamma	gamma	shape, scale=1
geometric	geom	prob
hypergeometric	hyper	m, n, k
lognormal	lnorm	meanlog=0, sdlog=1

名稱	R 函數	選項
logistic	logis	location=0, scale=1
multinomial	multinom	size, prob
normal	norm	mean=0, sd=1
negative binomial	nbinom	size, prob
Poisson	pois	lambda
Student's (t)	t	df
uniform	unif	min=0, max=1
Weibull	weibull	shape, scale=1
Wilcoxon's statistics	wilcox	m, n
	signrank	n

對於所給的分佈名稱，有四類。

以 func 為例，四類函數的對應為:
1. 「d」概率密度函數: dfunc(x, p1, p2, …), x 為數值向量;
1. 「p」(累積) 分佈函數: pfunc(q, p1, p2, …), q 為數值向量;
1. 「q」分位數函數: qfunc(p, p1, p2, …), p 為由概率構成的向量;
1. 「r」隨機數函數: rfunc(n, p1, p2, …), n 為生成數據的個數

這四類函數的第一個參數是有規律的: 形為 dfunc 的函數為 x，pfunc 的函數為 q，qfunc 的函數為 p，rfunc 的函數為 n

note: (但 rhyper 和 rwilcox 是特例，他們的第一個參數為 nn). 非中心參數 (non-centrality parameter) 僅對 CDF 和少數其它幾個函數有效.

$$\frac{\sum\limits_{i=1}^{n} x_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \sim N(\mu, \sigma^2/n)$$

uniform a~b

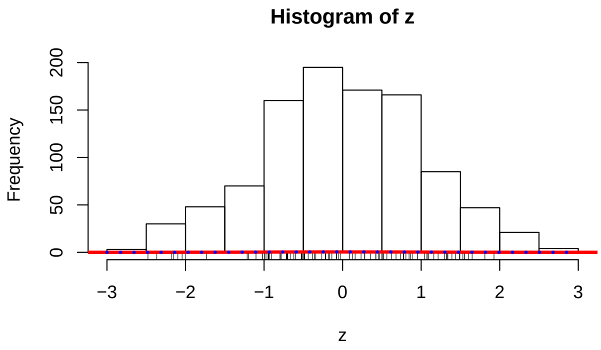
$$\mu = (a+b)/2$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$

data 的每一個 ROW 有 sample size (=i=column)
共 1000 次 (=N=row)

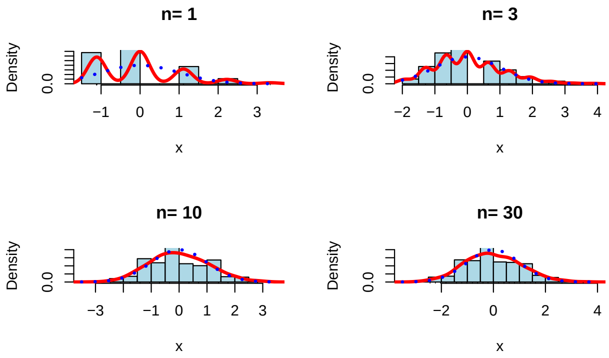
```
N=1000
i=3 #sample size
mu=0.5
sigma=1/sqrt(12)
```

```
data<-matrix(runif(i*N),ncol=i)
rs<-rowSums(data)
rs<-rs/i
z<-(rs-mu)/(sigma/sqrt(i))
hist(z)
lines(density(z), col = 'red', lwd = 3)
x<-z
curve(dnorm(x), col = 'blue', lwd = 3, lty = 3, add = T)
rug(sample(z,100))
```



```
limite.central <-
  function (r = runif, distpar = c(0, 1), m = .5, s = 1 / sqrt(12), n = c(1, 3, 10, 30
    for (i in n) {
      if (length(distpar) == 2) {
        x <-matrix(r(i * N, distpar[1], distpar[2]), nc = i)
      } else {
        x <-matrix(r(i * N, distpar), nc = i)
      }
      x <-(apply(x, 1, sum) - i * m) / (sqrt(i) * s)
      hist(x, col = 'light blue', probability = T, main = paste("n=", i),
        ylim = c(0, max(.4, density(x) $y)))
      lines(density(x), col = 'red', lwd = 3)
      curve(dnorm(x), col = 'blue', lwd = 3, lty = 3, add = T)
      if (N > 100) {
        rug(sample(x, 100))
      } else {
        rug(x)
      }
    }
  }
op <- par(mfrow=c(2,2))
```

```
limite.central(rbinom, distpar=c(10 ,0.1), m=1, s=0.9)
par(op)
```



n= 1

n= 3

n= 10

n= 30

- apply(x,1,sum) 第 2 個參數 1 表示 row 方向, 如果是 2 表示 column 和 matlab 相反。

```
需要安裝 babynames,ggplot2
```

Chapter 11

Tidy Basic 01

```
require(tidyr)
require(dplyr) # data_frame
```

From http://stackoverflow.com/questions/1181060

```
stocks <- data_frame(
  time = as.Date('2009-01-01') + 0:9,
  X = rnorm(10, 0, 1),
  Y = rnorm(10, 0, 2),
  Z = rnorm(10, 0, 4)
)
dset1 <- head(stocks)
knitr::kable(dset1, format = "html")
```

time

X

Y

Z

2009-01-01

-1.400

-1.11

1.87

2009-01-02

0.255

1.26

1.45

2009-01-03

-2.437

4.13

-5.22

2009-01-04

-0.006

-3.26

2.95

2009-01-05

0.622

1.02

7.55

2009-01-06

1.148

-3.73

-0.39

```
gather(stocks, stock, price, -time)
```

```
# A tibble: 30 x 3
   time       stock    price
   <date>     <chr>    <dbl>
 1 2009-01-01 X     -1.40   
 2 2009-01-02 X      0.255  
 3 2009-01-03 X     -2.44   
 4 2009-01-04 X     -0.00557
 5 2009-01-05 X      0.622  
 6 2009-01-06 X      1.15   
 7 2009-01-07 X     -1.82   
 8 2009-01-08 X     -0.247  
 9 2009-01-09 X     -0.244  
10 2009-01-10 X     -0.283  
# ... with 20 more rows
```

```
stocks %>% gather(stock, price, -time)
```

```
# A tibble: 30 x 3
   time       stock    price
   <date>     <chr>    <dbl>
 1 2009-01-01 X     -1.40
 2 2009-01-02 X      0.255
 3 2009-01-03 X     -2.44
 4 2009-01-04 X     -0.00557
 5 2009-01-05 X      0.622
 6 2009-01-06 X      1.15
 7 2009-01-07 X     -1.82
 8 2009-01-08 X     -0.247
 9 2009-01-09 X     -0.244
10 2009-01-10 X     -0.283
# ... with 20 more rows
```

```
dset1 <- head(stocks)
knitr::kable(dset1, format = "html")
```

time

X

Y

Z

2009-01-01

-1.400

-1.11

1.87

2009-01-02

0.255

1.26

1.45

2009-01-03

-2.437

4.13

-5.22

2009-01-04

-0.006

-3.26

2.95

2009-01-05

0.622

1.02

7.55

2009-01-06

1.148

-3.73

-0.39

設定 css

```
writeLines("td, th { padding : 6px } th { background-color : brown ; color : white; bo
```

```
stocks <- data_frame(
  time = as.Date('2009-01-01') + 0:9,
  X = rnorm(10, 0, 1),
  Y = rnorm(10, 0, 2),
  Z = rnorm(10, 0, 4)
)
dset1 <- head(stocks)
knitr::kable(dset1, format = "html")
```

time

X

Y

Z

2009-01-01

0.935

0.140

3.448

2009-01-02

0.176

-1.278

-0.973

2009-01-03

0.244

-0.100

-0.824

2009-01-04

1.624

-0.503

0.077

2009-01-05

0.112

0.890

0.118

2009-01-06

-0.134

5.511

2.199

```r
demo<-gather(stocks, stock, price, -time)

dset1 <- head(demo)
knitr::kable(dset1, format = "html")
```

time

stock

price

2009-01-01

X

0.935

2009-01-02

X

0.176

2009-01-03

X

0.244

2009-01-04

X

1.624

2009-01-05

X

0.112

2009-01-06

X

-0.134

11.1 long and wide data

11.1.1 gather: wide to long

```
wide <- data_frame(
  time = as.Date('2009-01-01') + 0:9,
  X = rnorm(10, 0, 1),
  Y = rnorm(10, 0, 2),
  Z = rnorm(10, 0, 4)
)

long <- gather(wide,stock,price,-time)
head(long)
```

```
# A tibble: 6 x 3
  time       stock  price
  <date>     <chr>  <dbl>
1 2009-01-01 X      1.07
2 2009-01-02 X     -0.665
3 2009-01-03 X      1.11
4 2009-01-04 X     -0.246
5 2009-01-05 X     -1.18
6 2009-01-06 X     -0.976
```

11.1.2 spread :long to wide

```
wide2 <-spread(long,stock,price)
head(wide2)
```

```
# A tibble: 6 x 4
  time            X      Y     Z
  <date>      <dbl>  <dbl> <dbl>
1 2009-01-01  1.07  -2.94   7.70
2 2009-01-02 -0.665  0.568  5.19
3 2009-01-03  1.11   2.67   3.00
4 2009-01-04 -0.246  0.473  2.22
5 2009-01-05 -1.18   2.64  -2.19
6 2009-01-06 -0.976  1.05   4.44
```

更多參考: cookbook for R

11.2 dplyr

函數名功能

- row_number 排序, 如果數值一樣, 則靠前出現的元素排名在前, 例如 (3,3) 則 1,2
- min_rank 排序, 如果數值一樣, 則都是同一等級，但是，佔用下一名次。例如

```
data<-c(3,3,4)
data
```

```
[1] 3 3 4
```

```
min_rank(data)
```

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

- dense_rank 排序, 如果數值一樣, 則都是同一等級，但是，不佔用下一名次

```
data<-c(3,3,4)
data
```

```
[1] 3 3 4
```

```
dense_rank(data)
```

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

- percent_rank 按百分比的排名
  percent_rank = (min_rank(x) - 1)/(sum(!is.na(x)) - 1)

- cume_dist 累計分佈
- ntile：floor(n * (row_number(x) - 1)/len + 1)

```
data<-round(runif(10)*10)
pr<-percent_rank(data)
cd<-cume_dist(data)
mr<-min_rank(data)
df<-data.frame(data,pr,mr,cd)
arrange(df,data)
```

```
   data    pr mr  cd
1     1 0.000  1 0.3
2     1 0.000  1 0.3
3     1 0.000  1 0.3
4     2 0.333  4 0.5
5     2 0.333  4 0.5
6     5 0.556  6 0.6
7     7 0.667  7 0.8
8     7 0.667  7 0.8
9     9 0.889  9 1.0
10    9 0.889  9 1.0
```

note:

(2,3,3,3,3,4,5,6,6,9)

1	2	3	4	5	6	7	8	9
0	1	4	1	1	2	0	0	1

arrange(dataframe, col1, col2, col3)

vs. dataframe[order(dataframe$col1, dataframe$col2, dataframe$col3),]
vs. with(dataframe, dataframe[order(col1, col2, col3),])

想要由大到小, 例如分數等級

```
data
```

```
 [1] 7 1 9 1 9 1 5 2 7 2
```

```
row_number(desc(data))
```

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

Percentile The nth percentile of an observation variable is the value that cuts off the first n percent of the data values when it is sorted in ascending order.

Problem Find the 32nd, 57th and 98th percentiles of runiform(200).

```
data<-runif(200)
quantile(data, c(.32, .57, .98))
```

11.3 other

```
library(babynames)
babynames
```

```
# A tibble: 1,858,689 x 5
    year sex   name          n   prop
   <dbl> <chr> <chr>     <int>  <dbl>
 1  1880 F     Mary       7065 0.0724
 2  1880 F     Anna       2604 0.0267
 3  1880 F     Emma       2003 0.0205
 4  1880 F     Elizabeth  1939 0.0199
 5  1880 F     Minnie     1746 0.0179
 6  1880 F     Margaret   1578 0.0162
 7  1880 F     Ida        1472 0.0151
 8  1880 F     Alice      1414 0.0145
 9  1880 F     Bertha     1320 0.0135
10  1880 F     Sarah      1288 0.0132
# ... with 1,858,679 more rows
```

11.3.1 Basic verbs

```
babynames %>% select(-prop)
```

```
# A tibble: 1,858,689 x 4
    year sex   name          n
   <dbl> <chr> <chr>     <int>
 1  1880 F     Mary       7065
 2  1880 F     Anna       2604
 3  1880 F     Emma       2003
 4  1880 F     Elizabeth  1939
 5  1880 F     Minnie     1746
 6  1880 F     Margaret   1578
 7  1880 F     Ida        1472
 8  1880 F     Alice      1414
 9  1880 F     Bertha     1320
10  1880 F     Sarah      1288
# ... with 1,858,679 more rows
```

```
babynames %>% select(year:n)
```

```
# A tibble: 1,858,689 x 4
    year sex   name          n
   <dbl> <chr> <chr>     <int>
 1  1880 F     Mary       7065
 2  1880 F     Anna       2604
 3  1880 F     Emma       2003
 4  1880 F     Elizabeth  1939
 5  1880 F     Minnie     1746
 6  1880 F     Margaret   1578
 7  1880 F     Ida        1472
 8  1880 F     Alice      1414
 9  1880 F     Bertha     1320
10  1880 F     Sarah      1288
# ... with 1,858,679 more rows
```

```
# starts_with(), ends_with(), contains()

babynames %>% filter(name == "Hadley")
```

```
# A tibble: 159 x 5
    year sex   name        n      prop
   <dbl> <chr> <chr>   <int>     <dbl>
 1  1906 M     Hadley      6 0.0000416
 2  1908 M     Hadley     16 0.0000962
 3  1909 M     Hadley    14 0.0000792
 4  1910 M     Hadley     5 0.0000240
 5  1911 M     Hadley     9 0.0000373
 6  1912 M     Hadley    11 0.0000244
 7  1913 M     Hadley    10 0.0000186
 8  1914 M     Hadley    15 0.0000220
 9  1915 M     Hadley    14 0.0000159
10  1916 M     Hadley    14 0.0000152
# ... with 149 more rows
```

```
babynames %>% filter(year == 1900, sex == "F")
```

```
# A tibble: 2,225 x 5
    year sex   name          n   prop
   <dbl> <chr> <chr>     <int>  <dbl>
 1  1900 F     Mary      16707 0.0526
 2  1900 F     Helen      6343 0.0200
 3  1900 F     Anna       6114 0.0192
 4  1900 F     Margaret   5304 0.0167
 5  1900 F     Ruth       4765 0.0150
 6  1900 F     Elizabeth  4096 0.0129
 7  1900 F     Florence   3920 0.0123
 8  1900 F     Ethel      3896 0.0123
 9  1900 F     Marie      3856 0.0121
10  1900 F     Lillian    3414 0.0107
# ... with 2,215 more rows
```

```
babynames %>% filter(year == 2013, sex == "F")
```

```
# A tibble: 19,203 x 5
    year sex   name          n    prop
   <dbl> <chr> <chr>     <int>   <dbl>
 1  2013 F     Sophia    21171 0.0110
 2  2013 F     Emma      20905 0.0109
 3  2013 F     Olivia    18397 0.00958
 4  2013 F     Isabella  17599 0.00916
 5  2013 F     Ava       15225 0.00793
 6  2013 F     Mia       13127 0.00683
 7  2013 F     Emily     13107 0.00682
 8  2013 F     Abigail   12387 0.00645
 9  2013 F     Madison   10575 0.00551
10  2013 F     Elizabeth  9431 0.00491
# ... with 19,193 more rows
```

```
babynames %>%
  mutate(
    first = tolower(substr(name, 1, 1)),
    last = substr(name, nchar(name), nchar(name))
  )
```

```
# A tibble: 1,858,689 x 7
    year sex   name          n   prop first last
   <dbl> <chr> <chr>     <int>  <dbl> <chr> <chr>
 1  1880 F     Mary       7065 0.0724 m     y
 2  1880 F     Anna       2604 0.0267 a     a
 3  1880 F     Emma       2003 0.0205 e     a
 4  1880 F     Elizabeth  1939 0.0199 e     h
 5  1880 F     Minnie     1746 0.0179 m     e
 6  1880 F     Margaret   1578 0.0162 m     t
 7  1880 F     Ida        1472 0.0151 i     a
 8  1880 F     Alice      1414 0.0145 a     e
 9  1880 F     Bertha     1320 0.0135 b     a
10  1880 F     Sarah      1288 0.0132 s     h
# ... with 1,858,679 more rows
```

```
babynames %>%
  arrange(desc(prop))
```

```
# A tibble: 1,858,689 x 5
    year sex   name        n   prop
   <dbl> <chr> <chr>   <int>  <dbl>
 1  1880 M     John     9655 0.0815
 2  1881 M     John     8769 0.0810
 3  1880 M     William  9531 0.0805
 4  1883 M     John     8894 0.0791
 5  1881 M     William  8524 0.0787
 6  1882 M     John     9557 0.0783
 7  1884 M     John     9388 0.0765
 8  1882 M     William  9298 0.0762
 9  1886 M     John     9026 0.0758
10  1885 M     John     8756 0.0755
# ... with 1,858,679 more rows
```

```
babynames %>%
  summarise(n = sum(n))
```

```
# A tibble: 1 x 1
```

```
          n
      <int>
1 340851912
```

11.3.2 Group by

分組指令不會影響到原來的資料

```
head(mtcars)
```

```
                   mpg cyl disp  hp drat   wt qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.62 16.5  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.88 17.0  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.32 18.6  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.21 19.4  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.44 17.0  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.46 20.2  1  0    3    1
```

```
str(mtcars)
```

```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
by_cyl <- mtcars %>% group_by(cyl)
head(by_cyl)
```

```
# A tibble: 6 x 11
# Groups:   cyl [3]
    mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  21       6   160   110  3.9   2.62  16.5     0     1     4     4
2  21       6   160   110  3.9   2.88  17.0     0     1     4     4
3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
```

```
4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
```

```
str(by_cyl)
```

```
Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame': 32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
 - attr(*, "vars")= chr "cyl"
 - attr(*, "drop")= logi TRUE
 - attr(*, "indices")=List of 3
  ..$ : int  2 7 8 17 18 19 20 25 26 27 ...
  ..$ : int  0 1 3 5 9 10 29
  ..$ : int  4 6 11 12 13 14 15 16 21 22 ...
 - attr(*, "group_sizes")= int  11 7 14
 - attr(*, "biggest_group_size")= int 14
 - attr(*, "labels")='data.frame':  3 obs. of  1 variable:
  ..$ cyl: num  4 6 8
  ..- attr(*, "vars")= chr "cyl"
  ..- attr(*, "drop")= logi TRUE
```

但是分組結果會影響其他 dplyr 指令的計算結果:

```
by_cyl %>% summarise(
  disp = mean(disp),
  hp = mean(hp)
)
```

```
# A tibble: 3 x 3
    cyl  disp    hp
  <dbl> <dbl> <dbl>
1     4  105.  82.6
2     6  183. 122.
3     8  353. 209.
```

```
by_cyl %>% filter(disp == max(disp))
```

```
# A tibble: 3 x 11
# Groups:   cyl [3]
    mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
2  24.4     4  147.    62  3.69  3.19  20       1     0     4     2
3  10.4     8  472    205  2.93  5.25  18.0     0     0     3     4
```

11.3.3 summarize()

What other summary functions can we use inside the summarize() verb? Any function in R that takes a vector of values and returns just one. Here are just a few:

mean(): the mean AKA the average
sd(): the standard deviation, which is a measure of spread
min() and max(): the minimum and maximum values respectively
IQR(): Interquartile range
sum(): the sum
n(): a count of the number of rows/observations in each group. This particular summary function will make more sense when group_by() is covered in Section 5.5.

11.3.3.1 實驗 pipeline vs no pipeline

產生資料

```
year=c(1990,    1991,   1990,   1991,   1990,   1991,   1990,   1991,   1990,   1991)
sex=c("f",  "f",    "f",    "f",    "f",    "m",    "m",    "m",    "m",    "m")
#value=c(1, 2,  3,  4,  5,  1,  2,  3,  4,  5)
value=c(1,  2,  3,  4,  5,  6,  7,  8,  9,  10)
df<-data.frame(sex,year,value)
head(df)
```

```
  sex year value
1   f 1990     1
2   f 1991     2
3   f 1990     3
4   f 1991     4
5   f 1990     5
6   m 1991     6
```

11.3.3.2 不用 pipeline

```
df<-group_by(df,sex)
ndf<-mutate(df,rank=min_rank(value))
arrange(ndf,sex)
```

```
# A tibble: 10 x 4
# Groups:   sex [2]
   sex    year value  rank
   <fct> <dbl> <dbl> <int>
 1 f      1990     1     1
 2 f      1991     2     2
 3 f      1990     3     3
 4 f      1991     4     4
 5 f      1990     5     5
 6 m      1991     6     1
 7 m      1990     7     2
 8 m      1991     8     3
 9 m      1990     9     4
10 m      1991    10     5
```

11.3.3.3 使用 pipeline

```
ndf<-df %>%
  group_by(sex) %>%
  mutate(rank = min_rank(value))

arrange(ndf,sex)
```

```
# A tibble: 10 x 4
# Groups:   sex [2]
   sex    year value  rank
   <fct> <dbl> <dbl> <int>
 1 f      1990     1     1
 2 f      1991     2     2
 3 f      1990     3     3
 4 f      1991     4     4
 5 f      1990     5     5
 6 m      1991     6     1
 7 m      1990     7     2
 8 m      1991     8     3
 9 m      1990     9     4
10 m      1991    10     5
```

問題: 1. 如何知道 min-rank(value) 中的 value 是全局或是欄位? hint: rm(value) 2. 會出現甚麼結果

df %>% group_by(sex) %>%str()

```
nb <-babynames %>%
  group_by(name)
```

```
babynames %>%
  group_by(name) %>%
  summarise(n = sum(n))
```

```
# A tibble: 95,025 x 2
   name           n
   <chr>      <int>
 1 Aaban         87
 2 Aabha         28
 3 Aabid          5
 4 Aabriella     15
 5 Aada           5
 6 Aadam        218
 7 Aadan        122
 8 Aadarsh      173
 9 Aaden       4218
10 Aadesh        15
# ... with 95,015 more rows
```

```
babynames %>%
  filter(name %in% c("John", "Mary", "William")) %>%
  group_by(name, sex) %>%
  summarise(n = sum(n))
```

```
# A tibble: 6 x 3
# Groups:   name [?]
  name    sex         n
  <chr>   <chr>   <int>
1 John    F       21657
2 John    M     5095674
3 Mary    F     4118058
4 Mary    M       15158
5 William F       15911
6 William M     4071645
```

```
babynames %>%
  group_by(year, sex) %>%
  mutate(rank = min_rank(desc(n))) %>%
  tail()
```

```
# A tibble: 6 x 6
# Groups:   year, sex [1]
   year sex   name       n       prop  rank
  <dbl> <chr> <chr>  <int>      <dbl> <int>
1  2015 M     Zyah       5 0.00000247 12008
2  2015 M     Zykell     5 0.00000247 12008
3  2015 M     Zyking     5 0.00000247 12008
4  2015 M     Zykir      5 0.00000247 12008
5  2015 M     Zyrus      5 0.00000247 12008
6  2015 M     Zyus       5 0.00000247 12008
```

11.3.3.4 Combining to answer more complex questions

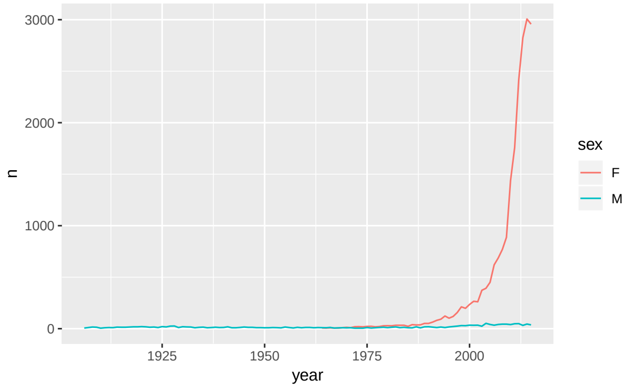
11.3.3.4.1 How many Hadley's?

```
babynames %>%
  filter(name == "Hadley") %>%
  group_by(sex) %>%
  summarise(n = sum(n))
```

```
# A tibble: 2 x 2
  sex       n
  <chr> <int>
1 F     21119
2 M      1814
```

11.3.3.4.2 The travesty

```
library(ggplot2)
babynames %>%
  filter(name == "Hadley") %>%
  ggplot(aes(year, n)) +
    geom_line(aes(colour = sex))
```


3000
2000
1000
0
n
1925
1950
1975
2000
year
sex
F
M

Chapter 12

programming in R

12.1 條件指令

12.1.1 if

一般來講，通用寫法如下：

```
if(判斷式){
 成功
}else{
 失敗
}
```

例如 # 多行寫法

```
if(4 > 3){
  TRUE
}else{
  FALSE
}
```

另外，如果區塊內只有一行

```
if(4 > 3) TRUE else FALSE
```

12.1.2 ifelse 的寫法

`ifelse('判斷式','成功','失敗')`
例如，`ifelse(4 > 3, T, F)`

12.1.3 switch 的寫法

switch(‘判斷式 ‘, ’ 動作 1’, ‘動作 2’, …) ** 指定第幾行 **

```
switch(2,
       1,
       2, # 因為判斷式為 2, 因此，執行動作 2
       3)
```

** 也可以利用字串 **

```
switch("a",
       a = 1,
       b = 2,
       c = "why")
```

```
[1] 1
```

這個 switch 並不是一個語法，而是一個函數，原理大概是第 1 個參數以後的東西，會被變成是一個 named list。例如上面的例子中

```
alist <-list(a=1,b=2,c="why")
```

然後利用判斷式查找要返回對應的值。

12.2 迴圈指令

R 語言的迴圈，可以搭配 break 和 next，前者跳出迴圈，後者繼續下一個迴圈。

12.2.1 for-loop

計算 1+2+3+4. 的值是多少?

```
rst <- 0
for(i in c(1:4)){
  rst <- rst + i
}
```

12.2.2 while-loop

```
i <- 1
result <- 0

while(i < 5){
  rst <- rst + i
  i <- i + 1
}
```

12.2.3 repeat-loop

```
i <- 1
rst <- 0

repeat{

  if(i > 4) break
  rst <- rst + i
  i <- i + 1
}
```

12.3 範例

12.3.1 範例 detach_package()

討論內建函數 `detach_package()`
使用方法:

```
detach_package(vegan)
detach_package("vegan", TRUE)
```

注意一下第 2 個使用方式，第 2 個參數 ‘character.only’:

```
detach_package <- function(pkg, character.only = FALSE)
{
  if(!character.only) # 如果沒有指定參數是字串
  {
    pkg <- deparse(substitute(pkg)) # 轉字串
  }
  search_item <- paste("package", pkg, sep = ":")
  while(search_item %in% search())
  {
    detach(search_item, unload = TRUE, character.only = TRUE)
  }
}
```

12.3.2 範例 2：.libPaths()

```
.libPaths
```

```
function (new)
{
    if (!missing(new)) {
        new <- Sys.glob(path.expand(new))
        paths <- c(new, .Library.site, .Library)
        paths <- paths[dir.exists(paths)]
        .lib.loc <<- unique(normalizePath(paths, "/"))
    }
    else .lib.loc
}
<bytecode: 0x55a96ae19f38>
<environment: 0x55a96ae1b898>
```

12.4 R 語言的 meta 計算。

- assign()
- 語法解析：substitute(), parse(), deparse()

- 表達式構造：quote()
- 表達式求值：eval(), source()
- 表達式：expression()

```
parse(text='x<-1+2')
```

```
expression(x <- 1 + 2)
```

在 R 中解析有三種不同的變種：

- The read-eval-print loop
- Parsing of text files
- Parsing of character strings

read-eval-print 是讀入文本，然後進行解析，然後求值，最後打印，這個就是我們日常看到的命令行操作。

解析文件是 parse 函數來完成。

解析字符串就是 parse 中用 text 參數來表示。

12.4.1 deparse

相對於 parse,deparse 把 `expression` 轉為字串。

```
deparse(expression(1+2))
```

```
[1] "expression(1 + 2)"
```

12.4.2 quote

quote() 的參數也是 `expression`，返回類別 `class`，可以再給 `eval()` 計算。`quote()` 的參數雖然是 `expression` 但是不會進行計算，而是 token 分解。

```
quote(1+2)
```

```
1 + 2
```

```
typeof(quote(1+2))
```

```
[1] "language"
```

```
class(quote(1+2))
```

```
[1] "call"
```

12.4.3 eval

12.4.3.1 parse() 和 eval()

text –>parse()–> expression—>eval() #### quote() 和 eval() expression –>quote()–> call—>eval()

```
eval(quote(1+2))
```

```
[1] 3
```

一般是用 parse 從字符串（或者是硬盤上的文件）解析成一個 expression 對象，是表達式列表，不是一個表達式。

```
eval(parse(text="1+2"))
```

```
[1] 3
```

首先使用 parse() 函數將字符串轉化為表達式（expression），而後使用 eval() 函數對表達式求解。

```
?eval
# Evaluate an R expression in a specified environment.
# Usage
# eval(expr, envir = parent.frame(),
#      enclos = if(is.list(envir) || is.pairlist(envir))
#          parent.frame() else baseenv())
# evalq(expr, envir, enclos)
# eval.parent(expr, n = 1)
# local(expr, envir = new.env())
```

12.4.3.2 substitue()

如果沒有第二個參數，是否 substitue()==quote()?

```
(s.e <- substitute(expression(a + b), list(a = 1)))   #> expression(1 + b)
```

```
expression(1 + b)
```

```
(s.s <- substitute( a + b,            list(a = 1)))   #> 1 + b
```

```
1 + b
```

12.5 討論

12.5.1 測試紀錄

```
test1<-function (..., list = character() )
{
    names <- c(as.character(substitute(list(...))[-1L]), list)
    return (names)
}
test1(x,y,list="data")
```

```
[1] "x"    "y"    "data"
```

`names <- c(as.character(substitute(list(...))[-1L]), list)` * list(…) 把傳進來的參數打包成 list * 在執行 list(…) 的時候，如果傳進的參數是還沒有初值的變數，會出錯，因此用 `substitute()` 打包。+ 為什麼用 substitute() 就不會出錯，因為它不執行這個 expression(e.g. list(…))。-為什麼可以不執行? 因為 `substitue()` 有把它變成匿名函數 (和 call 這個類別有關，自推論)，這樣當然不會有問題。-為什麼有-1L? 接續之前的推論，第一個元素應該是匿名函數的內建名稱。

所以整個函數，就是把使用者所有的參數，打成字串向量。如範例結果。

12.6 練習

1. zeros(3,2),ones(2,3)

hint:

```
sapply(a,function(x) matrix(x,2,2), simplify='array')
```

```
zeros<-function (nrow,ncol)
{
rst<-sapply(c(0),function(x) matrix(x,nrow,ncol),simplify='array')
  return(rst[,,1])
}
x<-zeros(4,5)
x
```

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

Chapter 13

Econometrics

13.1 package recommended

package `foreign` 可以用來導入外部資料例如，* stata 的資料 dta, * SPSS * SAS * Systat * Mini tab . mtp

利用 foreign 讀入 stata 的資料檔:

```
require(foreign)
affairs<- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/affairs.dta")
save(affairs,file='./resources/affairs.RData')
```

測試

```
rm(affairs)
load('./resources/affairs.RData')
head(affairs)
```

```
  id male age yrsmarr kids relig educ occup ratemarr naffairs affair
1  4    1  37    10.0    0     3   18     7        4        0      0
2  5    0  27     4.0    0     4   14     6        4        0      0
3  6    1  27     1.5    0     3   18     4        4        3      1
4 11    0  32    15.0    1     1   12     1        4        0      0
5 12    0  27     4.0    1     3   17     1        5        3      1
6 16    1  57    15.0    1     5   18     6        5        0      0
  vryhap hapavg avgmarr unhap vryrel smerel slghtrel notrel
1      0      1       0     0      0      0        1      0
2      0      1       0     0      0      1        0      0
3      0      1       0     0      0      0        1      0
```

```
4      0      1       0     0      0      0        0      0
5      1      0       0     0      0      0        1      0
6      1      0       0     0      1      0        0      0
```

Chapter 14

rbook 提要

14.1 index.Rmd

通常是一系列文章裡面的第一個文件。其他 rmd 文件的 yml 頭部都可以統一放在這裡。

下面是一些選項, 可以自行添加到上面:

```
bibliography: [book.bib, packages.bib]
biblio-style: apalike
link-citations: yes
github-repo: rstudio/bookdown-demo
```

bibliography 暫時沒有參考文件檔, 所以我沒有放.

14.2

- HTML Format 來自rmarkdown from rstudio
- pandox

14.3 其他 _output.yml 範例

簡單講,_output.yml 放的是輸出到哪裡?gitbook? pdf? 或者是簡單 html 格式。

14.4 _output.yml 範例

```
bookdown::gitbook:
  css: style.css
  config:
    toc:
      before: |
        <li><a href="./">A Minimal Book Example</a></li>
      after: |
        <li><a href="https://github.com/rstudio/bookdown" target="blank">Published wit
    edit: https://github.com/rstudio/bookdown-demo/edit/master/%s
    download: ["pdf", "epub"]

# bookdown::pdf_book:
#   latex_engine: "xelatex"
#   keep_tex: true
#   includes:
#     in_header: header.tex

 bookdown::pdf_book:
   includes:
     in_header: latex/preamble.tex
     before_body: latex/before_body.tex
     after_body: latex/after_body.tex
   keep_tex: true
   dev: "cairo_pdf"
   latex_engine: xelatex
   citation_package: natbib
   pandoc_args: ["--top-level-division=chapter", "--lua-filter=latex/sidebar.lua"]
   template: null
   quote_footer: null
   toc_unnumbered: false
   number_sections: true
```

```
---
output:
  html_document:
    highlight: tango
    number_sections: yes
    toc: yes
---
```

上面這段 YML 會被解譯到下面的 R 程式碼。猜測:html_document 對應 rmarkdown::html_document

例如, number_sections: yes 會被當成 `html_document()` 的參數轉譯:

highlight: tango 對應到參數 (沒有在下面)

```
rmarkdown::html_document( theme= "lumen",
                              template= template,
                              css= css,
                              toc= toc,
                              toc_float= TRUE,
                              toc_depth= 2,
                              number_sections= number_sections,
                              df_print = "paged",
                              code_folding = code_folding,
                              includes = includes(before_body = header)
                           )
}
```

14.5 .travis.yml

rmarkdown and travis

應該是用來指導怎麼建立專案的一個 yml。已知的部分有: directories : 可以指定子目錄, 用來放暫存檔案 (一些翻譯過程產生的中間黨), 否則每個 RMD 都會產生一個組目錄。

```
language: R
sudo: false
cache:
  packages: true
  directories:
  - _bookdown_files
  - $HOME/.npm

pandoc_version: 2.1.1
```

14.6 紀錄

before_chapter_script: “common.R” 前面要有兩個空白

if(FALSE) { ’ 兩個地方可以放, 1) 一個是在 _bookdown.yml 中的最後一行: before_chapter_script: “common.R” 2) 一個是在每個 RMD 的第一個 chunk 執行 source(“common.R”)

’ }

14.6.1 knitr hook

14.6.1.1 顏色

設定警告訊息為紅色

```
knitr::knit_hooks$set(error = function(x, options) {
  paste0("<pre style=\"color: red;\"><code>", x, "</code></pre>")
})
```

其他範例

```
#Color Format
colFmt = function(x,color){
  outputFormat = opts_knit$get("rmarkdown.pandoc.to")
  if(outputFormat == 'latex')
    paste("\\textcolor{",color,"}{",x,"}",sep="")
  else if(outputFormat == 'html')
    paste("<font color='",color,"'>",x,"</font>",sep="")
  else
    x
}
```

有關 hook 的設定, 參考 knitr hook

14.6.2 terminal output

問題: 如果以要看檔案內容, 而不想被無關的輸出符號干擾, 例如 [1], 那麼如果再 R SCRIPT 中可以使用 `system("cat 檔案名稱")` , 但是在 R Markdown 中 (如下) 不會顯示出結果:

```
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE, na="")
system("cat MyData.csv")
```

You can use either

{r, engine=‘bash’, comment=’‘} cat ’resources/hw.csv’

or

```
cat(readLines('resources/wh.csv'), sep = '\n')
```

```
# this is a test file
height,weight,sex
156,56, m
167,., f
189,70, m
180,, f
```

14.7 python

First you need to set the knitr options.

{r} knitr::opts_chunk$set(engine.path = list(python = '/anaconda/bin/python')) @@@ From that point on it just works.

{python} import this @@@ –>

14.8 產生 PDF 文件

14.8.1 Overview

參考

整個轉換過程如下：Rmarkdown -> markdown ->pandoc->pdf,html

14.8.2 meta data 上的 pdf 設定，常見版本

```
---
title: "Habits"
author: lendjwjcen
date: 3/1/2017
output: pdf_document
---
```

14.8.2.1 加入目錄

```
---
title: "Habits"
output:
  pdf_document:
    toc: true
    toc_depth: 2
---
```

TOC 深度預設是 3。yaml 的註解符號是 `#` 。

如果要章節號碼, 則利用 `number_sections option`:

```
---
title: "Habits"
output:
  pdf_document:
    toc: true
    number_sections: true
---
```

14.8.2.2 圖形選項

fig_width 和 fig_height 控制圖形的寬度和高度 (6 x 4.5 is used by default)

fig_crop 控制 pdfcrop utility (如果有) 是否自動使用來修剪圖型 (default). 如果我們的圖形設備是 postscript, 建議關掉這個選項。.

fig_caption 控制是否有圖型標題 (預設是否)

dev 控制用來描繪圖型的設備 (defaults to pdf)

例如:

```
---
title: "Habits"
output:
  pdf_document:
    fig_width: 7
    fig_height: 6
    fig_caption: true
---
```

14.8.2.3 Data Frame Printing

`df-print` 參數可以用來改進 data frame 的輸出, 選項包括:

選項描述 default: 利用 print.data.frame 函數
kable 其實利用的是 knitr::kable 函數
tibble 則利用 tibble::print.tbl_df 函數。.
例如:

```
---
title: "Habits"
output:
  pdf_document:
    df_print: kable
---
```

14.8.2.4 Syntax Highlighting

打光型態有 “default”, “tango”, “pygments”, “kate”, “monochrome”, “espresso”, “zenburn”, and “haddock” 如果不要打光, 則設定 null:

例如:

```
---
title: "Habits"
output:
  pdf_document:
    highlight: tango
---
```

14.8.2.5 LaTeX Options

有關 LaTeX 選項的設定位置並不是在 `output: pdf_document` 的後續段落, 而是在最頂層。例如:

```
---
title: "Crop Analysis Q3 2013"
output: pdf_document
fontsize: 11pt
geometry: margin=1in
---
```

14.8.2.6 其他

Available metadata variables include:

Variable	Description
papersize	paper size, e.g. `letter`, `A4`
lang	Document language code
fontsize	Font size (e.g. 10pt, 11pt, 12pt)
documentclass	LaTeX document class (e.g. article)
classoption	Option for documentclass (e.g. oneside); may be repeated
geometry	Options for geometry class (e.g. margin=1in); may be repeated
linkcolor, urlcolor, citecolor	Color for internal, external, and citation links (red, green, magenta, cyan, blue, black)
thanks	specifies contents of acknowledgments footnote after document title.

更多控制可以在 here. 找到

14.8.3 LaTeX Packages for Citations

citation 的處理, 預設是經由 `pandoc-citeproc`(不僅 PDF 文件, 還有 HTML), 但是對於 PDF 來講, 最好還是使用 LaTeX 來處理例如 `natbib` or `biblatex`。設定方法僅僅是設定 `citation_package` 為 `natbib` 或 `biblatex`, 例如:

```
---
output:
  pdf_document:
    citation_package: natbib
---
```

14.9 中文

建立中文 PDF 需要安裝 TeX 版本可以使用 `tinytex`。

```
install.packages('tinytex')
tinytex::install_tinytex()
```

14.10 Rstudio chinese

```
cd /usr/lib/rstudio/bin
sudo mkdir Qt
sudo mv libQt5* Qt
sudo mv qt.conf Qt
```

14.10.1 LaTeX Engine

RStudio 預設使用 `pdflatex`. 我們可以使用 `latex_engine` 選項設定其他引擎例如, “pdflatex”, “xelatex”, and “lualatex”. 而中文的引擎範例:

```
---
title: "Habits"
output:
  pdf_document:
  latex_engine: xelatex
---
```

但是只有這樣還不夠，因為我們需要中文字型, 這時可以利用 `indclude` 選項: ### Include 要在文件本身 (body) 前面後面加上一些內容, 可以利用 `includes` 選項:

```
---
title: "Habits"
output:
  pdf_document:
    includes:
      in_header: header.tex
      before_body: doc_prefix.tex
      after_body: doc_suffix.tex
---
```

最後的中文設定可以是這樣

```
---
title: "Untitled"
author: "lin"
date: " "
output:
  pdf_document:
    latex_engine: xelatex
    includes:
      in_header: header.tex
```

`header.tex` 的內容 (% 是註解):

```
\usepackage{xeCJK}
% microsoft windows

% linux font
\setCJKmainfont{AR PL UMing TW}

% microsoft font
% \setCJKmainfont{微軟正黑體}
% \setCJKmainfont{Microsoft YaHei}

%\setmainfont{Georgia} % 設定英文字型
%\setromanfont{Georgia} % 字型
%\setmonofont{Courier New}
```

14.10.2 註記

有關一些中文常常發生的錯誤:

- 無法解決字型錯誤:Package fontspec Error: The font “Inconsolata” cannot be found 改用其他字形，例如使用指令

```
fc-list :lang=zh-tw fullname
```

新系統例如 linux 安裝完以後，如果要測試中文是否可用於產生 pdf，最好利用一些短文。### 測試中文轉 pdf 下列的檔案名稱為 `x1.tx`, 測試的時候，可以利用 `tinytex::xelatex(x1.tex)` 用來驗證是否中文可以轉 pdf。

這裡先利用 tex 小檔案測試 xelatex 是否功能正常。

```
\documentclass{book}
\usepackage{xeCJK}
\setCJKmainfont{AR PL UMing TW}


\begin{document}
早安，台灣

$$b^2-4\times a \times c$$

night TeXing.
\end{document}
```

14.10.3 Custom Templates

也可以利用選項 `template` 取代 pandoc 模板:

```
---
title: "Habits"
output:
  pdf_document:
    template: quarterly_report.tex
---
```

參考 pandoc templates 上的說明。或者參考 default LaTeX template 中的範例。

14.10.4 Pandoc Arguments

如果在 YAML 中找不到相關於 pandoc 的選項, 則可以直接利用 `pandoc_args`. 例如:

```
---
title: "Habits"
output:
  pdf_document:
    pandoc_args: [
      "--no-tex-ligatures"
    ]
---
```

14.11 共用設定

多個文件如果都有一樣的輸出格式, 則可以利用檔案 `_output.yaml` 例如:

_output.yaml

```
pdf_document:
  toc: true
  highlight: zenburn
```

做為文件中的 YAML 的比對, 可以發現, 只是把 output: 的後面整個層級搬到 _output.yaml。

```
---
title: "Habits"
output:
  pdf_document:
    includes:
      in_header: header.tex
      before_body: doc_prefix.tex
      after_body: doc_suffix.tex
---
```

14.12 討論

- hook
  Verbatim Chunks in R Markdow
  *HTML widgets
- R Notebook HTML FOrmat
- Render option hook example
  *more
  dynamic book

14.12.1 inline code

inline 讓內籤 R 程式碼不作用: 方法是兩邊打上 “。例如:

```
inline("dd")
===>  `` `r dd` ``
```

```
cat(readr::read_file("rmd_topic_1.Rmd"))
```

討論
* [hook](https://github.com/lmmx/devnotes/wiki/Rmarkdown-custom-knit-hook-to-compile-a
[Verbatim Chunks in R Markdow](https://rmarkdown.rstudio.com/articles_verbatim2.html)
*[HTML widgets](https://github.com/rstudio/rmarkdown-book/blob/master/16-widgets.Rmd)
* [R Notebook HTML FOrmat](https://rmarkdown.rstudio.com/r_notebook_format.html)
* [Render option](http://brooksandrew.github.io/simpleblog/articles/render-reports-dire
[hook example](https://selbydavid.com/2017/06/18/rmarkdown-alerts/)
*[more](https://gist.github.com/yihui/2629886)
[dynamic book](https://github.com/yihui/knitr-book)

### inline code
```{r , include = FALSE}
chunk <- "````"
inline <- function(x = "") paste0("`` `r ", x, "` ``")

library(tidyverse)
```

inline 讓內籤R程式碼不作用:方法是兩邊打上``。例如:
``` {r,eval=FALSE}
inline("dd")
===>  `` `r dd` ``
```

```{r,  comment = ""}
cat(readr::read_file("rmd_topic_1.Rmd"))
```

```{r}
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE, quote=F)
# system("cat myData.csv") # 這個不行
cat(readLines('MyData.csv'), sep = '\n')
cat(readr::read_file("MyData.csv"))
```

~~~
```
~~~

````
mtcas 中的資料筆數有 `r nrow(mtcars)`
````

```r
rst = read.csv('resources/wh.csv',comment.char="#",na.string='.')
write.csv(rst, file = "MyData.csv",row.names=FALSE, quote=F)
# system("cat myData.csv") # 這個不行
cat(readLines('MyData.csv'), sep = '\n')
```

```
height,weight,sex
156,56, m
167,NA, f
189,70, m
180,NA, f
```

```r
cat(readr::read_file("MyData.csv"))
```

```
height,weight,sex
156,56, m
167,NA, f
189,70, m
180,NA, f
```

mtcas 中的資料筆數有 32

Chapter 15

git note

在 ubuntu 中，`git push` 要避免一直輸入帳號密碼的方式為：1.

```
git config --global credential.helper 'store --file ~/.my-credentials'
```

2. 然後在 git push 的時候，輸入帳號和密碼（只要一次，就會被紀錄在上面指定的檔案，這裡的例子是,‘.my-credentials’。

Chapter 16

gitbook introduction

3 個地方需要注意: _output.yml _bookdown.yml index.rmd

16.1 _output.yml:

```
bookdown::gitbook:
  css: style.css
  split_by: chapter
  config:
    toc:
      collapse: subsection
      before: |
        <li><a href="./">A Minimal Bookdown Book</a></li>
      after: |
        <li><a href="https://github.com/rstudio/bookdown" target="blank">Published with bookdown<
bookdown::pdf_book:
  includes:
    in_header: preamble.tex
  latex_engine: xelatex
  citation_package: natbib
bookdown::epub_book:
  stylesheet: style.css
```

上面要修改的部分只有書名 (title of book)

16.2 _bookdown.yml:

```
book_filename: "bookdown-xx"
chapter_name: "Chapter "
repo: https://github.com/seankross/bookdown-start
output_dir: docs
rmd_files: ["index.Rmd", "01-Introduction.Rmd", "02-Diving-In.Rmd"]
clean: [packages.bib, bookdown.bbl]
new_session: yes
```

- 欄位 book_filename ： 書名 (PDF 或 EPUB) 例如本例的書名為 bookdown_xx.pdf。
- 欄位 chapter_name: 每個章節的前綴, 例如 01-Introduction.Rmd 第一個 H1 標籤為 # Introduction , 會變成 “Chapter 1 Introduction”。
- 欄位 repo field just designates a GitHub repository associated with this book but this is not a required field.
- 欄位 output_dir: HTML 檔案的輸出位置。同時也是 pdf 檔案的輸出位置。如果沒有設定這個欄位, 那麼預設輸出位置是 _book/ 。
- 欄位 rmd_files: 這是選擇性的, 如果沒有設定, 那麼專案子目錄下的所有 RMD 都會被 rendered。
- 欄位 new_session: 這也是選擇性的。如果是 new_session: yes 那麼每個 RMD 都在新的 R 連結 (session) 描繪 (rendered)，否則在同一個 session.

16.3 new_session 注意事項

但是我注意到,new_session 設定為 yes 的時候,md 檔案會被留下, 而設定為 no 的時候, 則不會。

同時, 在 new_session=no 的時候, 可以指定子目錄中的 RMD 檔案。

例如

```
language:
  ui:
    chapter_name: "Chapter "
new_session: no
after_chapter_script: clear_vars_and_pkgs.R
rmd_files:
  - "index.Rmd"
  - "my_sub_dir/chapter1.Rmd"
  - "my_sub_dir/chapter2.Rmd"
```

注意: new_session=yes 的時候, 會產生 md 檔案, 如果上傳到 Github 會被 jekyll 解讀要避免這個情況發生, 推測兩個解法 1. (OK) 在 docs 中放入.nojkeyll 這個檔案 (在 bash 中執行指令 touch .nojkeyll)
1. (不確定) 在 _bookdown.yml 的欄位 after_chapter_script: 中指定執行指令殺掉所有 md 檔案

16.4 index.rmd

另外一個相關設定的地方是 index.rmd 。這個檔案用來設定書的 cover, 和前幾頁。因此 Preface 和簡介可以放在這個檔案。這個檔案的前幾行通常是有關 yaml 的一些設定, 例如

```
---
title: "XXX title"
author: "len jwj cen"
date: "2018-1-1"
site: bookdown::bookdown_site
documentclass: book
#bibliography: [book.bib]
#biblio-style: apalike
link-citations: yes
github-repo: seankross/bookdown-start
url: 'http\://seankross.com/bookdown-start/'
description: "gitbook's simple setup"
---
```

應該更改的有 title, author, date, github-repo, url, and description 欄位。其他設定有 cover-image: 圖檔位置。

site: bookdown::bookdown_site 有這行就不用設定 _site.yml。

16.5 執行

bookdown::render_book("index.Rmd")

16.6 加入:Travis

使用 Travis 產生書一需要 3 個檔案, 這三個黨要放在 github repo 的根目錄: 3 個檔案的前 2 個, 可以直接從這裡複製: ### .Rbuildignore

^.*.Rproj\$ $^{\text{.Rproj.user\$}}$ .travis.yml\$

16.6.1 .travis.yml:

```
language: r
cache: packages

script:
  - Rscript -e 'bookdown::render_book("index.rmd")'
```

16.6.2 DESCRIPTION

要讓 travis 誤認為 package 所以, 只是放著, 內容不管

```
Package: placeholder
Title: Does not matter.
Version: 0.0.1
Imports: bookdown
Remotes: rstudio/bookdown
```