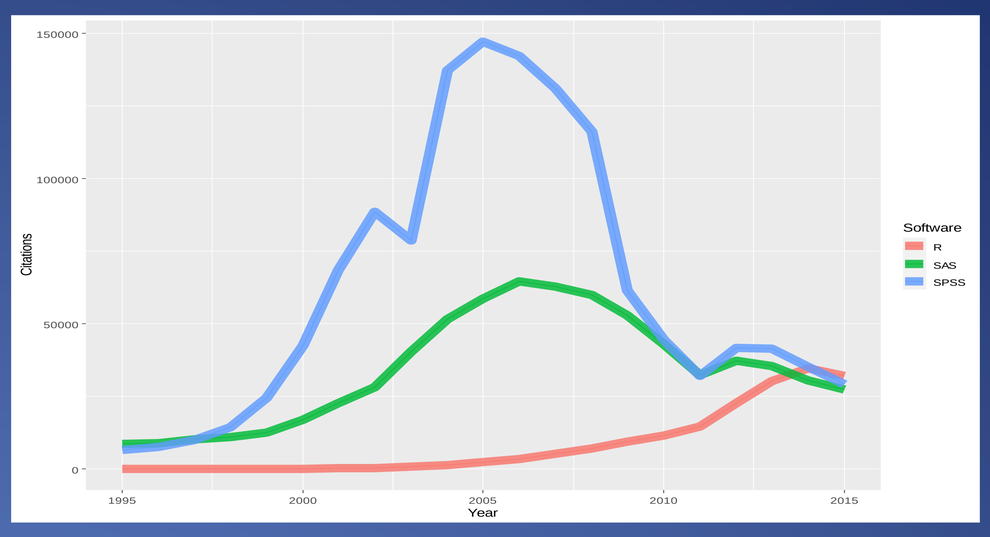
R與資料分析

10902 專題實作
職安2A 林書鳴

為什麼要使用R？

R是一種統計軟體，比起Python或C語言，R語言的語法較為獨特，但應用在統計分析上R語言將較勝一籌，下列網址將R 和其他統計軟體做比較，不難發現R支援了很多功能，自由度很高可以做許多運算和繪圖。

(https://en.wikipedia.org/wiki/Comparison_of_statistical_packages)



Estimated peer- 150000 reviewed publication citations for R, SAS and SPSS softwares

實際應用：

資料讀取

- 以read.table讀取各種外部檔案，透過sep =、header = TRUE、 as.is = TRUE 、comment.char =、skip = n等設定正確讀取正確格式數據。
- 如圖1中為以read.table("E:\\data\\file01.csv", sep=",", header=TRUE)，讀取在 E 磁碟資料夾 data 中的資料檔 file01.csv。

```
> rawdata_wide <- read.table(header=TRUE, text='
+ CaseID Gender Age Condition1 Condition2
+ 1 M 5 13 10.5
+ 2 F 6 16 11.2
+ 3 F 8 10 18.3
+ 4 M 9 9.5 18.1
+ 5 M 10 12.1 19
+ ')
> rawdata_wide
  CaseID Gender Age Condition1 Condition2
1      1      M   5       13.0       10.5
2      2      F   6       16.0       11.2
3      3      F   8       10.0       18.3
4      4      M   9        9.5       18.1
5      5      M  10       12.1       19.0
> rawdata_wide$subject <- factor(rawdata_wide$CaseID)
> rawdata_wide
  CaseID Gender Age Condition1 Condition2 subject
1      1      M   5       13.0       10.5       1
2      2      F   6       16.0       11.2       2
3      3      F   8       10.0       18.3       3
4      4      M   9        9.5       18.1       4
5      5      M  10       12.1       19.0       5
```

圖1. 以read.table()讀取各種外部檔案

資料檔新增變數與長寬格式轉換

- 使用在reshape2套件的 melt () 執行資料檔 exposure 長寬格式變換，將 day1 和 day2 的暴露測定值，轉換為長格式，並創造出兩個新變數 variable 和 value，標示原來 day1 和 day2 的訊息，如圖 2，並將資料檔儲存為 exposure01。
- 圖 3 為將圖 2 中於資料檔 exposure01 新增變數 variable 和 value，重新命名為 between 和 concentration 兩個變數，並將資料檔儲存為 exposure02。

```
> exposure
  idno     day1     day2 worker
1    1 1.914172 2.179576      1
2    2 4.471001 5.298556      2
3    3 1.801369 3.172514      1
4    4 4.279745 3.346465      2
5    5 4.358511 5.879843      1
6    6 4.780112 3.567759      2
> exposure01 <-melt(exposure, id=c("idno","worker"))
> exposure01
   idno worker variable    value
1     1      1     day1 1.914172
2     2      2     day1 4.471001
3     3      1     day1 1.801369
4     4      2     day1 4.279745
5     5      1     day1 4.358511
6     6      2     day1 4.780112
7     1      1     day2 2.179576
8     2      2     day2 5.298556
9     3      1     day2 3.172514
10    4      2     day2 3.346465
11    5      1     day2 5.879843
12    6      2     day2 3.567759
```

圖2. 以melt()於exposure新增between和value兩個變數

```
> exposure01
   idno worker variable    value
1     1      1     day1 1.914172
2     2      2     day1 4.471001
3     3      1     day1 1.801369
4     4      2     day1 4.279745
5     5      1     day1 4.358511
6     6      2     day1 4.780112
7     1      1     day2 2.179576
8     2      2     day2 5.298556
9     3      1     day2 3.172514
10    4      2     day2 3.346465
11    5      1     day2 5.879843
12    6      2     day2 3.567759
> exposure02 <-melt(exposure, id=c("idno","worker"),
+ variable.name=c("between"), value.name=c("concentration"))
> exposure02
   idno worker between concentration
1     1      1    day1      1.914172
2     2      2    day1      4.471001
3     3      1    day1      1.801369
4     4      2    day1      4.279745
5     5      1    day1      4.358511
6     6      2    day1      4.780112
7     1      1    day2      2.179576
8     2      2    day2      5.298556
```

圖3. 以melt()執行資料檔rawdata_wide長寬格式變換和新增變數

描述性統計值計算

- 以doBy套件執行資料檔 exposure02 中區分變數 between 的在 day1 和 day2 兩個類別 concentration 的描述性統計值的計算，結果包括：樣本數(concentration.n)、平均值(concentration.m)、標準差(concentration.s)、最大值(concentration.max)和最小值(concentration.min)，如圖4。

```
> install.packages("doBy")
Warning: package 'doBy' is in use and will not be installed
> library(doBy)
> summaryBy(concentration~between, data=exposure02,
+ FUN = function(x) { c(n=length(x), m = mean(x),
+ s = sd(x), max=max(x, na.rm=T), min=min(x, na.rm=T))})
  between concentration.n concentration.m concentration.s
1    day1              6        3.600818        1.361307
2    day2              6        3.907452        1.398663
  concentration.max concentration.min
1          4.780112          1.801369
2          5.879843          2.179576
>
>
```

圖4. 以doBy套件執行資料檔 exposure02 中between 和 concentration 的描述性統計

圖形描繪

- 以ggpubr套件描繪長條圖
- 使用資料檔 response01 描繪統計長條圖

```
> response01
           response number percentage
1    Disagree a lot    166       13.5
2 Disagree a little    116        9.4
3        No opinion    171       13.9
4    Agree a little    234       19.2
5       Agree a lot    542       44.1
> library(ggpubr)
> ggbarplot(response01, x="response", y="percentage", fill = "response",
+ color = "response", position = position_dodge(0.9), x.text.angle = 45) +
+ theme(legend.position="right") + labs(x="Response", y = "Percentage (%)")
```

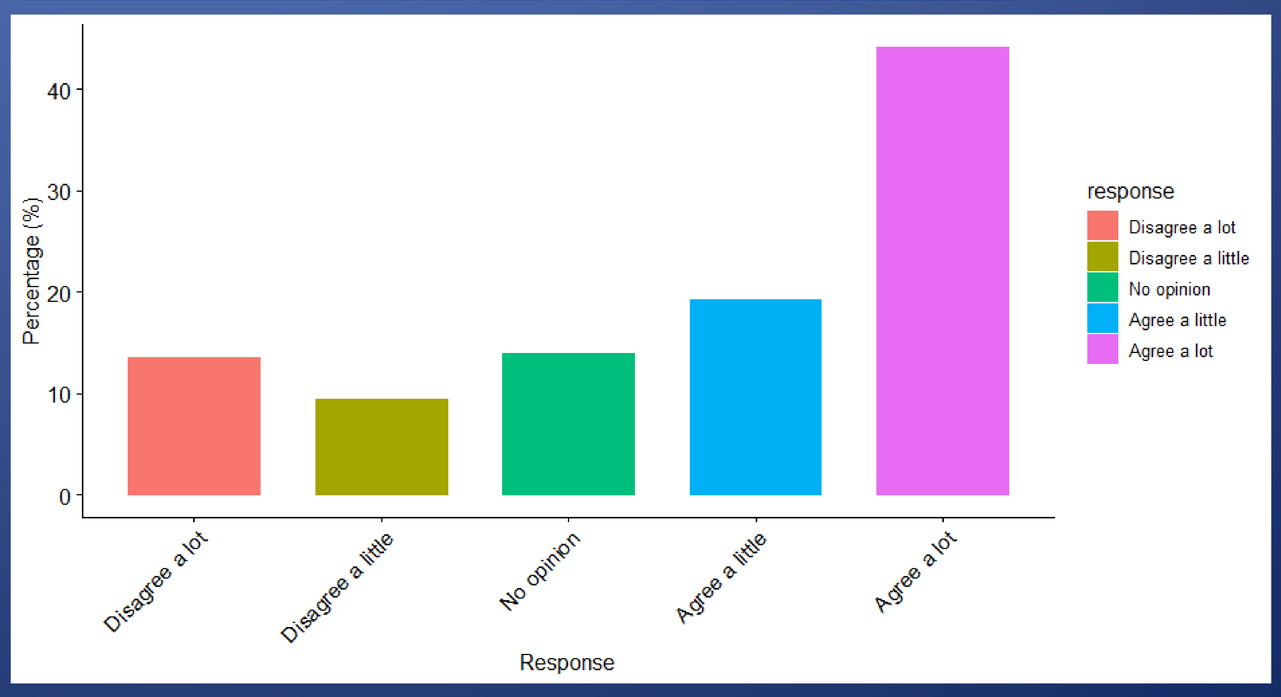


圖5. 以 ggpubr 描繪資料檔 response01 中變數 response 五個類別的百分比(percentage)