

R 練習題 (v2021.09)

吳漢銘*

國立政治大學統計學系



目錄

1 基礎: 物件、輸入輸出	2
2 程式設計	15
3 繪圖: Base graphics	48
4 繪圖: ggplot2	85
5 微積分、線性代數	99
6 機率與統計	111
7 資料處理	137
8 資料分析	147
9 其它	168

*hmwu@g.nccu.edu.tw

1 基礎：物件、輸入輸出

1.1 R 專案：

- (a) 建立一 R 專案，命名為「A 學號-姓名-R-exam1」(開頭首字為 A)。
- (b) 將程式碼檔「學號-姓名-R-exam1-Rcode.R」及答案卷檔「學號-姓名-R-exam1-Answer.txt」置於此專案中。

註：程式碼檔為撰寫本考題之 R 程式碼檔。答案卷檔的內容為程式在 console 執行之結果。

- (c) 下列指定題目作答完畢，將專案壓縮成「A 學號-姓名-R-exam1.zip」或「A 學號-姓名-R-exam1.rar」上傳。

1.2 假設作業系統是 MS Windows，並使用 Rgui 或 RStudio。

- (a) 印出目前 RStudio 工作區之系統資訊。
- (b) 印出系統現在之年月日及時間。
- (c) 列出目前目錄下的子目錄及檔案 (提示: `list.dirs`, `list.files`)。
- (d) 重新設定工作目錄在「C:\Users\Default」。
- (e) 印出目前 R 專案工作目錄及工作目錄下之所有目錄與檔案。
- (f) 印出目前 R 專案工作區 (Workspace) 的物件後，將之全部刪除。
- (g) 印出 R 套件安裝之目錄。
- (h) 印出現在電腦所使用的 R 軟體版本及 RStudio 軟體版本。
- (i) 以指令方式安裝套件「`seriation`」，安裝完，印出此套件的「版本」。
- (j) 於 R 提示符號 > 後，以指令方式查詢「&」這個符號的 help 說明。

1.3 安裝套件 (皆是以 R 指令完成)

- (a) 從台大資工 CRAN 鏡射站安裝兩個套件”`cluster`, `clValid`”，並載入 RStudio。
- (b) 到此位置<https://cran.r-project.org/web/packages/seriation/index.html> 下載 `seriation` 套件至電腦中。並在 Rgui 或 RStudio 中以 `install.packages` 指令安裝。
- (c) 在 Rgui 或 RStudio 中安裝三個 Bioconductor (<https://bioconductor.org>) 套件：`cancerclass`, `geneClassifiers`, `maSigPro`。
- (d) 印出電腦裡所有安裝的 R 套件。

1.4 列出電腦作業系統 (含位元數) 及 R 版本等等系統資訊。

1.5 (a) 用 `rep` 指令造出以下數列:

1 1 1 1 1 2 2 2 2 3 3 3 4 4 5

(b) 用 `rev` 和 `sequence` 指令造出以下數列:

1 2 3 4 5 6 2 3 4 5 6 3 4 5 6 4 5 6 5 6 6

1.6 產生數列

(a) "a" "b" "c" "d" "e" "b" "c" "d" "e" "c" "d" "e" "d" "e" "e"

(b) 1 4 7 10 13 16 19

1.7 產生數列:

(a) 用 `rep` 指令造出以下數列:

"A" "A" "A" "A" "A" "B" "B" "B" "B" "C" "C" "C" "D" "D" "E"

(b) 用 `seq`, `c` 指令造出以下數列:

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

(c) 產生以下數列:

$1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \dots, \frac{1}{19}, -\frac{1}{20}$

(d) 1 2 3 4 5 6 2 3 4 5 6 3 4 5 6 4 5 6 5 6 6

(e) 產生以下文字數列 (提示: `month.abb`, `seq`):

"Jan" "Mar" "May" "Jul" "Sep" "Nov" "Feb" "Apr" "Jun" "Aug" "Oct" "Dec"

1.8 產生下列數列 (Hint: `rep`, `seq`, `rev`):

(a) 8 7 6 5 7 6 5 4 6 5 4 3 5 4 3 2 4 3 2 1

(b) 3 7 11 15 19 23 27 31 35 39

1.9 利用 `rep`, `seq` 指令輸出下列向量:

(a) 3 3 3 3 3 3 3 3 3 3 3

(b) 1 1 1 4 4

(c) 2 2 2 5 5 5

(d) 2 2 4 4 6 6 8 8 10 10 12 12 14 14 16 16 18 18 20 20

1.10 產生以下有規律的數列。(提示: 儘量使用 `rep`, `seq`)

(a) 1 3 5 7 ... 21

- (b) 1 10 100 ... 10^9
- (c) 0 1 1 2 2 2 3 3 3 3 4 4 4 4 4
- (d) 1 2 5 10 20 50 100 ... 5×10^4
- (e) b d f h j l n p r t v x z

1.11 令 `colors <- c("red", "yellow", "blue")`。利用 `paste` 指令輸出下列文字向量

- (a) "red flowers" "yellow flowers" "blue flowers"
- (b) "redflowers" "yellowflowers" "blueflowers"
- (c) "several reds" "several yellows" "several blues"
- (d) "I like red, yellow, blue"

1.12 下列 `mydata` 物件的資料類別為何？將 `mydata` 轉成「因子」類別 (class) 之物件，並印出此因子物件中每一類別有多少個數。

```
set.seed(12345)
n <- sample(5:20, 1)
mydata <- sample(letters, n, replace=T)
```

1.13 某學生分析空氣品質資料 `airquality` 之風速 (Wind) 與溫度 (Temp) 的關係，他採用迴歸分析及共變異數分析，步驟如下：

```
lm.obj <- lm(airquality$Wind ~ airquality$Temp)
lm.anova <- anova(lm.obj)
lm.summary <- summary(lm.obj)
```

- (a) 物件 `lm.anova` 是屬於何種類別，其儲存結構如何？
- (b) 物件 `lm.summary` 有哪一些屬性可供存取？試取出 R^2 值。(提示: `r.squared`)

1.14 `cars` 是 R 內建資料集之一，紀錄車子當下之時速 (speed) 及煞車所需之距離 (dist)。某生使用以下程式碼做迴歸分析

```
cars.lm <- lm(dist ~ speed, data=cars)
cars.lm.sm <- summary(cars.lm)
```

試問 `cars.lm.sm` 是何種類別之物件？有哪些屬性可供存取？請單獨印出 F-statistic 之值。

1.15 (a) 輸入以下矩陣並命名為 `my.mat`。

$$\begin{bmatrix} 1 & 5 & 8 \\ 7 & 0 & 6 \\ 3 & 2 & 9 \\ 10 & 4 & 11 \end{bmatrix}$$

- (b) 將資料的列 (row) 命名為 `no.1`, `no.2`, `no.3`, `no.4` · 將欄 (column) 命名為 `var.1`, `var.2`, `var.3`。
- (c) 將 `var.3` 排序後 (由小到大) · 把資料矩陣依 `var.3` 的大小來排序。

1.16 下列為數個家庭的背景資料 ("NA" 代表無觀察值):

Name	Wife	No. of Children	Child Ages
George	Mary	3	4, 7, 9
Aaron	Sue	2	2, 5
John	Nico	0	NA
Tom	NA	1	10
Barrett	NA	NA	NA
Colin	Cathy	2	4, NA

- (a) 以 R 表列 (list) 類別方式將上述資料儲存為一 `list` 變數, 命名為 `family`。
- (b) 請單獨列出男主人 Barrett 家庭所有的資訊。
- (c) 是否可將上述 `family` 轉為 `data.frame` 類別之物件。

1.17 `family` 物件以表列方式紀錄數個家庭的背景資料 · 請單獨列出男主人 Barrett 家庭所有的資訊。

```
family <- list(name=c("George", "Aaron", "John", "Tom", "Barrett", "Colin"),
wife=c("Mary", "Sue", "Nico", NA, NA, "Cathy"),
no.children=c(3, 2, 0, 1, 2, NA),
is.own.house=c(T, T, F, F, T, NA),
child.ages=list(c(4,7,9), c(2, 5), NA, 10, c(NA, 4), NA))
```

1.18 利用 `substr`, `paste` 指令將電話號碼"203/781-1255" 換成"(203)7811255"。

1.19 由螢幕輸入 (讀入 `scan`)2 個數字 (例如: 26, 87) · 印出其總和。

1.20 請將 (0, 1) 區間等分為 10 份子區間:

- (a) 印出每一子區間之左端點。
- (b) 印出每一子區間之右端點。
- (c) 印出每一子區間之中點。

1.21 造出以下之矩陣，使得其行位名為 C1 ~ C5 及列位名為 R1 ~ R4:

	C1	C2	C3	C4	C5
R1	1	3	5	7	9
R2	2	4	6	8	10
R3	11	13	15	17	19
R4	12	14	16	18	20

1.22 有一學歷調查的資料 (degree) 如下:

```
set.seed(12345)
edu <- c("國小", "國中", "高中", "大學", "碩士以上")
degree <- sample(edu, 100, replace=T)
```

- (a) degree 是什麼 R 類別之物件?
- (b) 將此 degree 轉成 R 「因子類別」之物件，同時使其 levels 是依照學歷而排序。
- (c) 各學歷的人數有多少人?
- (d) 學歷為「高中」(含) 以上的人數有多少人?

1.23 下列 Letters.code 為一個包含「A」~「E」的向量。

```
set.seed(123456789)
Letters.code <- sample(LETTERS[1:5], 20, replace=T)
```

- (a) 將 Letters.code 中的「A」與「B」編為「第 1 組」，「C」編為「第 2 組」，「D」與「E」編為「第 3 組」。
- (b) 將上小題所得到的編組 Group.code，與 Letters.code 造成一個資料框 (data.frame)，使其具有 Letters.code 和 Group.code 兩欄位，且為順序之因子類別。印出此資料之內容及結構。(順序為 A<B<C<D<E; 第 1 組 < 第 2 組 < 第 3 組)

1.24 李克特量表 (Likert Scale) 的五等測量法是根據陳述語句的傾向給予各等級不同分數。對正向陳述而言，答案越正向分數越高例如：「非常同意」為 5 分，「同意」

為 4 分，「普通」為 3 分，「不同意」為 2 分，「非常不同意」為 1 分)。今有一問卷資料集，為一群學生(男生、女生)對某門課教師評分是否公平的認同程度(consent)，如下：

```
set.seed(1234567)
n <- 60
ID <- sample(1:n) #座號
gender <- sample(c("男", "女"), n, replace=T) # 性別
consent <- sample(c("非常不同意", "不同意", "普通", "同意", "非常同意"),
                  n, replace=T, prob=c(0.1, 0.1, 0.2, 0.4, 0.3)) # 認同程度
```

試回答下列各問題：

- 將性別轉為 R 因子 (factor) 類別。男女生各多少人？
- 將認同程度轉為 R 有順序的因子 (factor) 類別。印出
- 將上述資料存成一 R 資料框 (data.frame)，命名為 survey.df，欄位名稱依序為座號、性別及認同程度。印出此資料的前 5 筆紀錄。印出此資料框的結構。
- 列出填寫「不同意」(含) 以下的學生座號，共有幾人。(註：使用「<= 或 >=」)
- 此次調查結果，認同程度平均為多少分？

1.25 「statlog_vehicle_846x18.txt」是以 tab 為分隔的資料，具有 18 個變數，請讀入 R 之後，列出資料框維度、前後各 5 筆紀錄及儲存此資料框物件所佔用的記憶體。(原始資料說明：[https://archive.ics.uci.edu/ml/datasets/Statlog+\(Vehicle+Silhouettes\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes)))。)

1.26 有一班級學生之數學成績如下(成績已依座號順序排列)：

```
set.seed(12345)
math.score <- sample(0:100, 100, replace=TRUE)
```

- 計算前 10 位同學(座號 1 號至 10 號)之成績平均數。
- 成績及格之同學座號為何？共有多少人及格？
- 印出此資料的第一個四分位數。(提示：summary)(限單獨印出第一個四分位數之數值)

1.27 (a) 請將下列某班三科成績，以資料框 (data.frame) 之類別儲存(命名為 my.score) (需有欄位名稱)。同時，將此資料框的每一列(同學)的 ID 命名為 s1, s2, ..., s50。(b) 列出三科成績皆不及格之同學之 ID 及其成績。(NA 表示此位同學在某科缺考，以零分計算) (c) 將此三科成績以表列 (list) 類別之物件表示。

```
set.seed(12345)
score <- c(NA, 0:100)
math <- sample(score, 50, replace=T)
english <- sample(score, 50, replace=T)
chinese <- sample(score, 50, replace=T)
```

1.28 有一班級學生之數學成績如下 (成績已依座號順序排列):

43 94 20 8 46 72 93 8 28 33 79 60 93 52 8

- (a) 將資料輸入 R，並存至一向量物件，命名為 `math.score`。
- (b) 此成績紀錄，共有多少位同學？
- (c) 列印出偶數座號同學之成績，並計算其平均數。
- (d) 成績及格 (大於或等於 60 分) 之同學座號為何？共有多少人及格？

1.29 某班學生之成績和性別紀錄如下 (資料是依照學生座號 1、2、… 依序紀錄; NA 代表缺考):

成績: 30, 49, 95, NA, 54, NA, 61, 85, 51, 22, 0, 0 性別: m, f, f, m, f, m, m, f, f, m

- (a) 本班共有多少學生？男女生各多少人？
- (b) 此科目成績最高分及最低分是幾分？
- (c) 計算此科目成績平均及標準差。男女生成績平均各是多少？
- (d) 老師欲將成績依序做以下調整: (i) 缺考以 0 分計;(ii) 每人加 10 分 (缺考者不加分，超過 100 分以 100 分計)。印出調整後的分數。
- (e) 以調整後的分數計，列出及格 (60 分以上，含) 同學的座號。共有幾位？

1.30 有一班級 100 位學生之微積分期中及期末成績如下 (成績已依座號順序排列):

set.seed(12345) Calculus.midterm <- sample(0:100, 100, replace=TRUE) Calculus.final <- sample(0:100, 100, replace=TRUE)

- (a) 微積分期中成績之平均數及變異數為何？
- (b) 兩次成績之 (皮爾森) 相關係數為何？
- (c) 若每位同學的學期總成績為他們的期中及期末成績之平均，請問微積分總成績及格之同學座號為何？共有多少人及格？

- (d) 列出成績進步達 10 以上的同學之座號。(成績進步: 期末成績高於期中成績)
- (e) 奇數座號之同學，其微積分期末考不及格有幾人？
- (f) 列出兩次成績皆不及格的同學座號。(提示: 「 & 」)

1.31 某班「R 程式設計」一科學期各項成績總表紀錄於「R-score.xlsx」。

- (a) 讀取資料檔，印出前 5 位同學成績紀錄。
- (b) 計算各項考試(不含點名)平均分數及標準差。
- (c) 依照各項考試配分(小考 1(10%), 小考 2(15%), 小考 3(15%), 作業 (20%), 期末考 (40%)) 計算每位同學之學期成績，並以 `data.frame` 的類別型式印出學號及學期成績。(其它項目不用列出)

1.32 有一資料，紀錄某班級 10 位學生之姓名 (`student`)，組別 (`group`) 及分數 (`score`) 如下：

```
set.seed(12345)
group <- sample(letters[1:3], 10, replace=T)
score <- sample(0:100, 10, replace=T)
student <- c("Bruckner", "Caringer", "Mendoza", "Jaleela", "Williams",
           "Rida", "Kai", "Jaabir", "Garces", "Trevor")
```

- (a) 請將上述資料建立一 R 資料框 (`data.frame`) 物件(命名 `myData`) 如下：

```
> myData
  group score
Bruckner     c    80
Caringer     c    59
Mendoza      b    91
Jaleela       b    88
Williams      b   100
Rida          a    90
Kai            b    88
Jaabir        a    19
Garces        c    33
Trevor        a    78
```

- (b) 請將資料依組別及分數排序如下：(提示: `? order`)

	group	score
Jaabir	a	19
Trevor	a	78
Rida	a	90
Jaleela	b	88
Kai	b	88
Mendoza	b	91
Williams	b	100
Garces	c	33
Caringer	c	59
Bruckner	c	80

1.33 某三班 (Class) 同學之數學及英文考試成績如下:

```
Student: Bruckner, Caringer, Mendoza, Jaleela, Williams, Rida, Kai, Jaabir,
Garces, Trevor
Class: C, A, A, C, B, B, C, C, B, A
Math: 45, 33, 97, 71, 65, 39, 70, 54, 22, 48
English: 79, 26, 99, 76, 98, 22, 95, 15, 60, 95
```

- (a) 請將上述資料建立一 R 資料框 (data.frame) 類別之物件 (命名 Class.Score) · 使得學生姓名 (Student) 為此資料框之列位名; 欄位名則為班別 (Class) 及兩科目之分數 (Math、English)。印出 Class.Score。
- (b) 產生一個邏輯向量變數 (Pass) · 其中 TRUE 代表兩科目之平均分數有大於或等於 60 分。並將 Pass 合併至 Class.Score 中。印出 Class.Score。

1.34 讀取「stock-data.txt」資料檔，印出資料前 5 筆紀錄、後 5 筆紀錄。檢查 (印出) 資料每一變數 (欄位) 是否有符合 R 的類別物件，若沒有，請更改。(提示: 成交筆數、成交金額、成交股數皆為數值變數，不是字元變數)

1.35 讀取下列檔案，列印出資料前 5 筆，及後 5 筆紀錄；同時檢查 (印出) 資料每一變數 (欄位) 是否有符合 R 的類別物件，若沒有，請更改。[\(http://www.hmwu.idv.tw/web/R/data/\)](http://www.hmwu.idv.tw/web/R/data/)

- (a) R-score.xlsx
- (b) 20140714-weather.txt
- (c) weather_delays14.csv

1.36 mydata.xlsx 為某班之成績紀錄檔:

- (a) 利用 `read_excel {readxl}` 讀取檔案”mydata.xlsx” 的「calculus」工作表。印出資料前後各 5 筆紀錄。
- (b) 將上述讀入資料的欄位名稱重新命名為「No、Department、ID、Name、Quiz1、Quiz2、Quiz3、Quiz4、TA、MidCore1 MidCore2、MidSum」，並將「ID」指定為列位名稱。(因此資料就沒有「ID」這欄位了。)
- (c) 現各次考試的配分重新指定如下: 4 次小考，Quiz(1)~ Quiz(4)，各佔 8%，期中考 (MidSum) 佔 30%。期中總成績為上述各次考試之結算，若缺考以零分計，請以資料框方式，列出期中總成績不足 30 分之同學姓名 (Name) 及期中總成績 (含學號 (ID))，老師要寄發期中預警單。

1.37 有一成績資料檔「106-1-DA-Score.xlsx」，請依照資料檔內各次考試配分 (出席不計)，計算總成績，並將總成績附加在原成績資料表最後一欄後，存出成一 Excel 檔案: 「106-1-DA-Score_Final.xlsx」。

1.38 於網址<https://data.gov.tw/dataset/132344>, 下載資料:「臺北市家暴通報案件數統計資訊」，其儲存檔名為「106-108 家防中心報表.csv」。(註: 不得更改下載資料檔之內容。)

- (a) 於 RStudio，讀入此資料，印出前 10 筆紀錄及後 10 筆紀錄。
- (b) 將變數「間」轉成 R 的日期類別，並印出其結構 (`str`)。
- (c) 將變數「年齡區間」轉成是有順序的 R 因子類別，並印出其結構 (`str`)。
- (d) 選取變數「間」為”2019-7”(含) 之後的資料，並將此子資料集寫出 (匯出) 成一 Excel 檔案，命名為「2019_Violence.xls」。

1.39 資料來源: <https://github.com/owid/covid-19-data/tree/master/public/data>

- (a) 讀取「新冠肺炎」資料 (檔案: `owid-covid-data.csv`)，並印出前後各 5 筆紀錄。
- (b) 分別計算並印出三個國家 (Germany，United Kingdom 及 United States)，在本年度 8 月份「平均」新增確診案例 (`new_cases`)。

1.40 資料檔 `SalaryGov_Month_subset.xlsx` 為政府薪情平臺 <https://earnings.dgbas.gov.tw/experience.aspx> 匯出之「每人每月總薪資 (新臺幣元)」資料子集合。其中每一「行業類別」下，含有「性別」欄位。讀入資料檔，印出資料摘要 (`summary`) 及結構 (`str`)。請確認每一變數 (欄位) 皆是正確的 R 類別 (例如: 數值變數、日期變數)。若不是請做必要的轉換。

1.41 有一級數 S_n 如下:

$$S_n = \sum_{i=1}^n \frac{(-1)^{i+1}}{2i-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + \frac{(-1)^{n+1}}{2n-1}.$$

已知 $\lim_{n \rightarrow \infty} S_n = \pi/4$.

- (a) 產生 $(-1)^{i+1}, i = 1, \dots, n$ (其中 $n = 10$) 如右: 1 -1 1 -1 1 ... -1
- (b) 產生 $2i - 1, i = 1, \dots, n$ (其中 $n = 10$) 如右: 1 3 5 7 ... -19
- (c) 產生 $(-1)^{i+1}/(2i - 1), i = 1, \dots, n$ (其中 $n = 10$) 如右: 1 -1/3 1/5 -1/7 ... 1/19
- (d) 計算 $4S_{10}$, $4S_{100}$ 和 $4S_{1000}$ 。

1.42 有一函數 $f(x) = x^2 + x - 1$, 其定義域為 $(0, 1)$ · 請將 $(0, 1)$ 等分為 10 份子區間 · 將每一子區間之右端點所形成的集合稱為 x · 請計算 x 之函數值 $f(x)$ 。

1.43 一個人的肌肉質量預期會隨著年齡增長而下降。為探究這女性群體中的這個關聯性 · 一營養學家由年齡 40 到 79 歲的婦女中 · 每 10 歲一組隨機抽取 15 人進行研究。結果資料如「musclemass.csv」 · 其中 y 表肌肉質量 · x 表年齡。讀入資料 · 並列印出資料前 5 筆及後 5 筆紀錄。

1.44 某銷售人員在 2018 年的網路銷售紀錄從公司資料庫隨機抽樣 10 筆如下:

```
Dates: 0924, 1112, 1231, 1105, 0604, 0219, 0416, 0611, 0813, 1029
Time: 01:00, 04:00, 16:00, 23:00, 08:00, 09:00, 07:00, 17:00, 03:00, 14:00
Items: shirt, shirt, pants, jacket, jacket, shirt, jacket, jacket, shoes, shirt
Volume: 7951, 159, 1958, 6848, 3762, 3678, 8696, 9045, 6208, 1425
```

- (a) 請將上述資料儲存成一資料框 (data.frame) 類別之物件 · 命名 mySale · 使得第一個欄位為銷售日期時間 (DateTime) · 類別為 POSIXct, 時區為世界協調時間 (UTC); 第二個欄位為銷售品項 (Items), 類別為 factor; 第三個欄位為銷售量 (Volume), 類別為 numeric。印出 mySale。
- (b) 本資料中 · 七月 (含) 之後的銷售品項為何? 其總銷售量為多少?

1.45 資料 my.months 為某公司產品一年內的銷售月份 (1~12) 紀錄 · 以阿拉伯數字登記。

```
set.seed(12345)
my.months <- sample(1:12, 50, replace = TRUE)
```

- (a) 請將此阿拉伯數字登記之資料轉成英文簡寫月份 (命名為 `my.months.eng`):
 "Sep" "Nov" "Oct" ... "Jan" "Aug"。(提示: `month.abb`)
- (b) (承上小題) 各月份之銷售次數為何?
- (c) (承上小題) 下半年 (7~12 月) 之總銷售次數為何?
- (d) (承 (a) 小題) 請將此資料轉成一個依英文簡寫月份為順序的因子向量 (命名為 `my.months.eng.f`)。
- (e) (承上小題) 五月 (May) 至八月 (Aug) 的之總銷售次數為何?

1.46 某公司之銷售紀錄資料檔 `sales` 中，兩欄位資訊分別為某產品之銷售日期時間 (`date.time`) 及其銷售量 (`items.quantity`)。請計算此產品在 2015 年之平均銷售量。

```
my.time <- strptime(c("08/01/2014 00:00:00", "12/31/2018 23:59:59"), "%m/%d/%Y %H:%M:%S")
set.seed(12345)
date.time <- sample(seq(from = my.time[1], to = my.time[2], by='hour'), 100,
replace = T)
items.quantity <- sample(0:1000, 100, replace = T)
sales <- data.frame(date.time, items.quantity)
```

1.47 四群學生，人數 (`number`) 是 10、20、30、40 人，平均體重 (`weight`) 分別是 50、55、60、65 公斤，計算全部學生的平均體重。(提示: 將 `number` 及 `weight` 設定為 數字向量。)

1.48 某社區之 10 位住戶，接受體能量測之指數及滿意度調查資料如下 (NA 表示未接受量測或調查):

```
年紀(age): 54 64 75 21 66 49 25 72 50 72
性別(gender): "女" "男" "男" "女" "女" "男" "男" "女" "男" "女"
指數(index): 86 30 NA 43 35 42 31 7 29 80
滿意度(sat): "滿意" "非常滿意" "非常不滿意" "非常滿意" "普通" "非常不滿意" "普通" "滿意"
"普通" "非常滿意"
```

- (a) 將此資料輸入 R 中，共計 4 個變數: `age`, `gender`, `index`, 及 `sat`。將「滿意度 (sat)」設置成一個具有順序的因子類別之物件。(大至小的順序為「非常滿意」至「非常不滿意」)
- (b) 滿意度為「滿意」(含) 以上程度的人數共多少人。
- (c) 請計算年紀大於 40 歲男性之平均體能指數。

1.49 於網址<https://data.gov.tw/dataset/60139>, 下載資料:「臺北市公眾區免費無線上網熱點資料 (新版)」，儲存檔名為「Taipei_Free_AP.xlsx」。(註: 不得更改下載資料檔之內容。)

-
- (a) 於 RStudio，讀入此資料之前 10 筆紀錄 (命名為 TPE.wifi)。
 - (b) 由 TPE.wifi 選取「熱點名稱 (NAME)、緯度 (LATITUDE) 及經度 (LONGITUDE)」三欄位的資料，存成為一子資料集，命名為 TPE.wifi.subset，並印出此子資料集。
- 1.50 於網址<https://data.gov.tw/dataset/61797>, 下載資料：「臺北捷運全系統旅運量統計 _201803」，並儲存檔名為「臺北捷運全系統旅運量統計 _201803.csv」。(註：可更改下載資料檔之內容格式，但不得更改資料之正確性。)
- (a) 於 RStudio，讀入此資料 (命名為 TPE.MRT)，並直接列印前 3 筆及後 3 筆紀錄。
 - (b) 檢查 (印出) 資料每一變數 (欄位) 是否有符合 R 的類別物件 (日期)，若沒有，請更改。
 - (c) 選取日期 107/3/12~107/3/18 之資料 (需利用運算子 <, <=, >, >=)，計算此週之「總運量」平均數。

2 程式設計

2.1 丟 3 顆公平的骰子，其和為 `dice.sum`，

```
dice.sum <- sum(sample(1:6, 3, replace = TRUE))
```

試寫一 R 函式，印出總和 `dice.sum` 並做如下判別：如果和大於 13 點，則印出「厲害！」，反之印出「再加油！」。

2.2 (a) 請利用 `for` 寫一函式，計算一數列之平均數及變異數。

(b) 若有一成績紀錄如下

```
x <- sample(0:100, 50),
```

請利用上小題之函式算出平均數及變異數。

(c) 請與 `mean` 和 `sd` 之結果相比較。

2.3 利用 `for` 寫一函式，印出九九乘法表。

2.4 利用雙迴圈 `for`，印出下列圖形。

(a)

```
1
1 2
1 2 3
1 2 3 4
1 2 3 4 5
```

(b)

```
1
333
55555
7777777
999999999
```

2.5 輸入任何一個正整數 $n(n \leq 10)$ ，輸出 n 階層的 Pascal 三角形。

(例) 輸入: 5

輸出:

```
1
1   1
1   2   1
1   3   3   1
1   4   6   4   1
```

2.6 利用 `for`，試計算 $(1 \times 2 \times \dots \times 1000000)$ 之結果所需要的電腦系統時間。

2.7 (a) 計算 $n!$ 的程式可採用 (1) `for`, (2) `repeat`, (3) `while`, (4) 遞迴法及 (5) R 指令 `factorial`。(詳細程式見講義)。請用以上五種方法分別計算 $1000!$ 所需要的系統時間。

(b) 呈上題，請用指令 `system.time` 再分別計算一次。

2.8 設 $a_n = \frac{n+3}{n+8}$, $b_n = \frac{2n^2+3}{2n^2+8n}$, $c_n = \frac{\sqrt{n}}{2+\sqrt{n}}$, $n \geq 1$, 依定義可得 $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} c_n = 1$ ，請列出下表。

	n	an	bn	cn
1	1	0.4444444	0.5000000	0.3333333
2	10	0.7222222	0.7250000	0.6125741
3	20	0.8214286	0.8364583	0.6909830
4	30	0.8684211	0.8838235	0.7325211
5	40	0.8958333	0.9099432	0.7597469
6	50	0.9137931	0.9264815	0.7795188
7	60	0.9264706	0.9378906	0.7947869
8	70	0.9358974	0.9462355	0.8070727
9	80	0.9431818	0.9526042	0.8172560
10	90	0.9489796	0.9576241	0.8258876
11	100	0.9537037	0.9616827	0.8333333
12	10000	0.9995004	0.9996002	0.9803922

2.9 R 物件 `number` 是一個具有 1000 個正整數的數字向量，其數字範圍為 0 到 100，而數字於向量中的位置記為 1~1000。

```
set.seed(12345)
number <- sample(0:100, 1000, replace=T)
```

- (a) 使用 `for`，找出 `number` 數字向量中，第 100 個偶數的出現的位置，其數字為何。
- (b) 使用 `repeat`，找出 `number` 數字向量中，第 100 個偶數的出現的位置，其數字為何。
- (c) 使用 `while`，找出 `number` 數字向量中，第 100 個偶數的出現的位置，其數字為何。

2.10 某商業公司舉行抽獎活動，中獎名單紀錄於 `award-list.xlsx` 檔中，包含會員姓名、會員卡號及得獎金額。2.1 讀取資料

- (a) 請讀取此檔案，並印出全部中獎名單。

- (b) 因考量個資法，公告名單不能將全名公開，請你幫此名單，每一中獎者的姓名及會員卡號，部份字元打上 *，例如第一筆紀錄為「沈俞予 7113235607」，請改為「沈 * 予 7113***607」，印出修改後可公告之名單。(提示: substr)

	會員姓名	會員卡號	得獎金額
1	沈*予	7113***607	500
2	簡*榕	8010***376	1000
3	徐*良	9010***896	2000
4	賴*茹	3010***872	1500
5	林*玲	5011***845	4500
6	吳*鳳	2592***839	1000
7	江*翰	3714***694	3000
8	葉*鴻	4012***657	2500
9	阮*全	3053***421	5000
10	黃*鈴	3317***422	3500

- (c) 承上小題，請將修改後之名單，依照「得獎金額」由多至少的順序，全部印出。

	會員姓名	會員卡號	得獎金額
1	阮*全	3053***421	5000
2	林*玲	5011***845	4500
3	黃*鈴	3317***422	3500
4	江*翰	3714***694	3000
5	葉*鴻	4012***657	2500
6	徐*良	9010***896	2000
7	賴*茹	3010***872	1500
8	簡*榕	8010***376	1000
9	吳*鳳	2592***839	1000
10	沈*予	7113***607	500

2.11 有一 50 筆成績資料如下

```
score <- sample(0:100, 50, replace = TRUE)
```

判別此資料中是否有高於 95 分的同學，若有，印出「老師請同學吃飯」，若沒有印出「老師很生氣」。

2.12 某班學生期中考微積分及線代的成績資料如下：

```
set.seed(12345)
student.id <- paste("student", 1:50, sep=".")
Calculus <- round(rnorm(length(student.id), mean=65, sd=10), 0)
LinearAlgebra <- sample(0:100, length(student.id), replace = TRUE)
```

- (a) 印出兩科成績皆在 85 分以上的學生 id。(Hint: which)
- (b) 印出兩科成績皆在 60 分以下的學生 id。(Hint: which)
- (c) 各科成績最高分及最低分分別是哪些學生? (Hint: max, min)

2.13 某班 60 名學生之統計學成績 (stat) 依座號順序為

```
set.seed(12345)
stat <- sample(0:100, 60, replace=TRUE)
stat[sample(1:60, 5)] <- NA
```

其中 NA 為缺考。(a) 請計算此次成績之平均數 (缺考不計入)。(b) 若 3 號同學之成績登錄錯誤，76 分更正為 47 分，求更正後的成績平均數 (缺考以 0 分計)。

2.14 有一 50 筆課業成績資料如下

```
score <- sample(0:100, 50, replace = TRUE)
```

大學生課業成績以 60 分為及格，以 100 分為滿分，而「開根號再乘以 10」是著名的成績調分方式，請寫一函式，輸入為某班學生某科之成績，回傳：(1) 分數調整前被當學生之比例，(2) 分數調整前最高之成績，(3) 分數調整後被當學生之比例，及(4) 分數調整後最高之成績。

2.15 某班之期中考各科成績表格 ScoreTable 如下：

```
no <- 65
student.id <- paste("student", 1:no, sep=".")
set.seed(12345)
gender <- factor(sample(c("f", "m"), no, replace = TRUE))
Calculus <- sample(0:100, no, replace = TRUE)
LinearAlgebra <- sample(0:100, no, replace = TRUE)
English <- sample(0:100, no, replace = TRUE)
ScoreTable <- data.frame(student.id, gender, Calculus, LinearAlgebra, English)
```

- (a) 印出三科成績皆及格的同學 (student.id)。
- (b) 印出男生中，三科成績皆高於各自科目平均分數的同學 (student.id)。

2.16 小銘老師有某班學生之期中考試及加分考試兩筆資料。

```
set.seed(12345)
n <- 50
midterm <- sample(0:100, n, replace = TRUE)
extra <- sample(0:100, n, replace = TRUE)
```

成績比例為期中考 (midterm) 佔 40%、加分考 (extra) 佔 60%。兩次考試結算成績 (100%) 若小於期中考成績，則最後結算成績以期中考計。試寫一 R 函式 (自訂函式中需使用 ifelse)，處理上述計算，並回傳 (1) 最後結算成績之平均數及變異數，及 (2) 最後期中考試被當之學生比例。

2.17 某班學生 (student.id) 期中考微積分及線代的成績資料如下：

```
set.seed(12345)
student.id <- paste("student", 1:50, sep=".")  
Calculus <- round(rnorm(length(student.id), mean=65, sd=10), 0)  
LinearAlgebra <- sample(0:100, length(student.id), replace = TRUE)
```

老師註解成績的方法如下：

- i. 兩科成績皆高於 85 以上 (含)，記為「佳」。
 - ii. 任一科成績低於 40 以下 (含)，記為「要加強」。
 - iii. 兩科成績皆低於 40 以下 (含)，記為「危險」。
- (a) 利用 `for` 寫一 R 函式，計算「佳」「要加強」「危險」各有多少位同學。
- (b) 同一函式裡，再印出「佳」及「危險」之學生座號 (id)。

2.18 某班某科原始成績如下: `orig.score <- sample(0:100, 55, replace = TRUE)`。老師為了日行一善，打算調整學期總成績 (final.score)，其計算方法有以下三種選擇

- i. 維持原始分數不調分，但高於 55 分，低於 60 分者，加至 60 分及格。
- ii. 「開根號再乘以 10」。
- iii. 調成學期總成績最後之平均為 65 分，但高於 100 分者以 100 計。

試寫一 R 函式，包含上述三種調分方式 (使用者執行程式時，可自由選擇其中一種調分方式)，計算 (1) 原始成績之平均數及變異數；(2) 學期總成績之平均數及變異數；(3) 最後被當之學生比例。

2.19 某班學生 (student.id) 修課 5 科成績資料，分別由各科老師提供如下：

```

student.id <- paste("student", 1:55, sep=".")  

set.seed(123)  

Calculus <- round(rnorm(length(student.id), mean=65, sd=10), 0)  

LinearAlgebra <- sample(0:100, length(student.id), replace = TRUE)  

BasicMath <- sample(0:100, length(student.id), replace = TRUE)  

Rprogramming <- sample(0:100, length(student.id), replace = TRUE)  

English <- sample(0:100, length(student.id), replace = TRUE)

```

- (a) 請將此各別資料轉成單一資料表格 (命名為 mydata) · 使得欄位名稱為科目名 · 列名稱為學生的座號 student.id · 並列印出前 3 位同學成績紀錄。
- (b) 請將資料依 LinearAlgebra 排序後 · 印出此科目最高分及最低分各 5 位同學的各科成績。
- (c) 若每科學分數皆為 3 學分 · 同時每科以 60 分為及格。請找出 1/2 的同學。

2.20 某班學生 (student.id) 某科期中考成績 (score) 資料如下:

```

student.id <- paste("student", 1:50, sep=".")  

p <- dnorm(seq(-3,3,length=101))  

my.p <- p/sum(p)  

set.seed(123456)  

score <-sample(0:100, length(student.id), replace = TRUE, prob=my.p)

```

大學生課業成績以 60 分為及格 · 以 100 分為滿分 · 請寫一 R 函式 · 以「開根號再乘以 10」為調分方式 · 輸入為某科之成績 · 回傳:

- (a) 分數調整前 · 不及格學生之比例。
- (b) 分數調整前 · 最高成績之學生座號。
- (c) 分數調整後 · 全班成績之平均數及標準差。

2.21 檔案 score02.csv 記錄某班的統計學期中和期未成績。

- (a) 讀入資料 (使其具有欄位名稱: 「學號、期中考、期末考」), 印出前 7 筆紀錄。
- (b) 將欄位名稱依序更改為: id, mid, final。
- (c) 印出期未成績比期中成績進步的同學 id。
- (d) 將期中及期未成績 · 各分成及格和不及格兩組 · 則會有四種狀況 (例如其中一種: 期中及格 · 但期末不及格)。印出四種狀況之人數。
- (e) 學期成績的計算方式為期中考和期末考的平均成績 · 請將資料依學期成績由高分至低份排序印出。

2.22 有某班學生之微積分成績明細紀錄於資料檔 (score.txt) 中，其中成績以 60 分為及格，100 分為滿分，成績空白以零分計。學期總成績計算方法如下：(i) 配分比例為：小考成績佔 40%(各次小考平均配分)、期中考佔 30%、期末考佔 30%；(ii) 小考成績刪除其中最低分一次。

- (a) 請讀入此資料 (命名為 Score) 使得欄位名稱為考試別，列名稱為學號。列印出前 5 筆學生各次成績紀錄。
- (b) 將此資料具有遺失值 (NA) 的成績改為零分。列印修改後的資料 (命名為 my.score) 前 5 筆學生各次成績紀錄。
- (c) 學生學號為 s0050 的小考成績中，最低分數為第幾次？刪除此次成績後，其小考平均分數為何？
- (d) 小考成績中，每位學生的最低分數為第幾次？
- (e) 刪除每位同學之最低分小考成績後，試計算每位同學小考平均成績，其平均數及變異數為何？
- (f) 依學期總成績計算方法，計算學期總成績，其平均數及變異數為何？
- (g) 試寫一 R 函式，輸入為成績資料 my.score 及學期總成績，輸出為以下資訊：

```
> score.print(my.score, total)
```

本學期考試摘要表

	小考1	小考2	小考3	小考4	小考5	小考6	期中考	期末考	學期成績
平均數	52.85	36.25	57.72	54.42	45.98	38.26	56.15	43.12	51.01
變異數	949.07	648.50	649.15	547.88	542.25	554.65	705.81	619.43	392.13

不及格人數比例： 67.79 %

2.23 有某班學生之微積分成績明細紀錄於資料檔 (score2015.txt) 中，其中成績以 60 分為及格，100 分為滿分，成績空白以零分計。學期總成績計算方法如下：(i) 配分比例為：小考成績佔 40%(各次小考平均配分)、期中考佔 25%、期末考佔 25%、助教實習課佔 10%，出席次數分數為額外加分，每出席一次，加 2 分 (滿分 18 分)；成績紀錄共 8 項。(ii) 小考成績刪除其中最低分一次。

- (a) 請讀入此資料 (命名為 Score) 使得欄位名稱為性別、姓名及考試別 (中英文皆可)，列名稱為學號。列印出前 5 筆學生各項成績的紀錄。
- (b) 計算並印出六項成績，其每一項成績的最高分、最低分、平均分數及其變異數。(遺失值不列入計算)

小考1.	小考2.	小考3.	小考4.	期中考.	期末考.
最高分.					
最低分.					
平均.					
變異數.					

- (c) 將此資料具有遺失值 (NA) 的成績改為零分。刪除每位同學之最低分小考成績後，計算並印出每位同學小考總得分。
- (d) 依學期總成績計算方法，計算並印出每位同學的學期總成績。(超過 100 分，以 100 分計)
- (e) 請問不及格人數為多少？被當的比例為何？男女生被當的比例各又如何？

提示：小考刪除最差一次之後的計分方式，舉例如下：若有三次小考分為 60, 30, 90。配分為 5%, 6%, 7%。原始得分為 $60*0.05 + 30*0.06 + 90*0.07 = 11.1$ 若刪除最差一次成績後，所得分數為： $(60*0.05 + 90*0.07)*(5+6+7)/(5+7) = 13.95$

- 2.24 寫一函式 (`my.test`)，輸入為一組學生成績 (`score`)，判別此資料，若「成績及格人數達半數以上 (含)，且有 90 分以上 (含) 之同學」則印出「本次成績不調分，平均為：xx.xx」否則印出「本次成績會調分，不及格比例為：xx.xx」。(小數點以下兩位)

```
> set.seed(123456)
> score <- sample(1:100, 50, T)
> my.test(score)
本次成績不調分，平均為： 55.78
>
> set.seed(123456)
> score <- sample(1:100, 150, T)
> my.test(score)
本次成績會調分，不及格比例為： 60.67 %
```

- 2.25 有某班學生之學期各科總成績紀錄於資料檔 (`score1032.txt`) 中，其中成績以 60 分為及格，100 分為滿分，成績空白以零分計。七門科目 (英文，統計學，軟體入門，保險精算，數值分析，語表，離散數學) 之學分數依序為 2, 4, 3, 3, 3, 2, 3。

- (a) 計算每位同學之學業平均成績。請印出座號 1~10 號同學之「座號及平均成績」。
(不需印出 80 位學生之結果)
- (b) 計算每位同學通過科目數。請印出座號 11~20 號同學之「座號及通過科目數」。
(不需印出 80 位學生之結果)
- (c) 列印出所有「二一」同學的座號、學號、姓名及其學業平均成績。
- (d) 計算每位同學總得學分數。請印出女同學之「座號及總得學分數」。
- (e) 請依照學業平均成績將學生分成三組：低分組 (50 分 (含) 以下)、均分組 (50~70 分) 及高分組 (70(含) 分以上)。請印出下表。

各組人數	男生人數	軟體入門平均	平均通過科目數.
低分組.			
均分組.			
高分組.			

2.26 某班某次考試之成績 (ScoreData) 如下 · (a) 試計算每人之平均分數 。(b) 若三科成績 (math · english · algebra) 計算平均之權重依序為 (0.5, 0.2, 0.3), 試計算每人之加權平均分數 。(提示: apply, mean, weighted.mean)

```
set.seed(123456789)
math <- sample(0:100, 10, replace=T);
english <- sample(0:100, 10, replace=T)
algebra <- sample(0:100, 10, replace=T);
ScoreData <- cbind(math, english, algebra)
```

2.27 某班某次考試之四科成績如下:

```
set.seed(123456789)
n <- 10
math <- sample(0:100, n, replace=T);
english <- sample(0:100, n, replace=T)
algebra <- sample(0:100, n, replace=T)
programming <- sample(0:100, n, replace=T)
```

若四科成績 (math, english, algebra, programming) 計算平均之權重依序為 (0.4, 0.2, 0.3, 0.1), 試計算每人之加權平均分數 · 並將全班成績依加權平均分數之高低排序 。(排名第 1 為加權平均分數最高者)

rank	math	english	algebra	programming	weighted.mean
1	...				
2	...				
...					

2.28 以下為某校學生名字及某科目成績:

```
student <- c("John", "Mary", "Tom", "George", "Berry", "Nico", "Tim", "Jessica", "David")
score <- c(70, 58, 87, 22, 94, 30, 69, 94, 60)
```

利用 which 指令 · 列出哪個學生成績最高 · 哪個學生成績最低, 哪些學生的成績在平均以下 。

2.29 (a) 讀入資料 score-data.txt(其類別為 data frame) · 命名為 my.score 物件 · 使得欄位名稱為科目名 · 列名稱為學號 。列印出前 5 位同學所有成績紀錄 。

(b) 將資料 `my.score` 的列 (row) 命名為 `student.1`, `student.2`, ..., `student.n`。(n 為 row 的個數)。

(c) 將「基數」的成績以「開根號乘以十」重新計算後，結合全班其它各科成績匯出成另一資料檔 `new-score.txt`，內容需有欄位名稱，列名稱，並以 TAB 作分隔，而且輸出資料不要有引號。

2.30 小翔是一個對未來充滿抱負的青年，在工作之餘仍不忘利用下班時間充實自己所學，他審視大環境的趨勢、工作的性質與自己的專長，決定利用下班補習英文 (X 小時) 與電腦 (Y 小時)，假設英文課程補習費每小時 400 元，電腦課程補習費每小時 600 元，而小翔一個月的進修預算 (Budget) 上限為 12,000 元，其效用函數為 $U = X^{1/2}Y^{1/2}$ ，試寫一個 R 函式 (命名為 `study`)，輸入為補習英文與電腦的時數及預算 (內定值為 12,000)。輸出為以下表格，其中 `Tuition` 為所需的學費，`U` 為效用函數值，`Fit` 為學費沒有超出預算之註記。

	Eng.hr	Comp.hr	Tuition	U	Fit
1	13	8	10000	10.19804	*
2	14	8	10400	10.58301	*
3	15	8	10800	10.95445	*
... (中間省略)					
23	15	12	13200	13.41641	
24	16	12	13600	13.85641	
25	17	12	14000	14.28286	

2.31 試寫一 R 函式 (命名為 `triangle_side_length`) 計算三角形之三邊長，其中輸入為直角座標系上之三個點座標 $(A(x_1, y_1), B(x_2, y_2), C(x_3, y_3))$ ，輸出為三個點座標所形成的三角形之三邊長。程式執行以 $A(3, 2), B(5, 8), C(12, 4)$ 三點座標為範例。

2.32 海龍公式 (Heron's formula 或 Hero's formula)，是利用三角形的三條邊長 (a, b, c) 來求取三角形面積 (A) 的一個方法，其公式如下

$$A = \sqrt{s(s - a)(s - b)(s - c)}, \quad \text{where } s = \frac{a + b + c}{2}.$$

試寫一 R 函式 (命名為 `Heron`) 計算三角形之面積，其中輸入為三角形之三邊長，輸出為此三角形之面積。程式執行，三邊長以 7、8、9 為範例。

2.33 有某一試卷之測驗結果，紀錄於"answer.txt"。試卷中 10 題選擇題之正確答案依序為

B, D, B, D, D, A, C, D, C, B

(a) 請讀取此資料，並列印前 5 筆紀錄。

```
> first5.records
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
s1 C D D A D A B C C B
s2 B D B D D A C D B B
s3 B A A B D A C B C B
s4 B D B A B C C D C B
s5 B D D D A C C D A B
```

(b) 若某學生之答案為

A, D, B, D, B, A, B, D, C, B

試問他答對哪些題目。若答對一題得 10 分，則此學生所得之總分為何？

```
> correct.item
[1] 2 3 4 6 8 9 10
> n.correct
[1] 70
```

提示: as.integer, as.factor, which

(c) 若答對一題得 10 分，請計算每個人的總得分，並印出得分表格如下：

```
> score.table
  0 10 20 30 40 50 60 70 80 90 100
  9 18 16 9 18 19 27 34 25 10 6
```

提示: t, apply, table

(d) 若設定總得分前 25% 為高分組，總得分後 25% 為低分組，則哪些學生是高分組，哪些學生是低分組，而人數各為多少人。

```

> rownames(answer)[topID]
[1] "s2"    "s12"   "s16"   "s19"   "s20"   "s21"   "s25"   "s31"
[9] "s38"   "s41"   "s43"   "s47"   "s52"   "s54"   "s66"   "s69"
[17] "s73"   "s79"   "s80"   "s86"   "s95"   "s96"   "s102"  "s112"
[25] "s128"  "s129"  "s139"  "s143"  "s146"  "s149"  "s153"  "s157"
[33] "s158"  "s164"  "s175"  "s176"  "s182"  "s184"  "s185"  "s188"
[41] "s190"

> rownames(answer)[lowID]
[1] "s17"   "s27"   "s35"   "s36"   "s37"   "s49"   "s56"   "s57"
[9] "s58"   "s64"   "s65"   "s71"   "s72"   "s81"   "s82"   "s83"
[17] "s87"   "s90"   "s93"   "s97"   "s105"  "s107"  "s108"  "s113"
[25] "s120"  "s123"  "s125"  "s131"  "s132"  "s134"  "s145"  "s148"
[33] "s161"  "s163"  "s165"  "s168"  "s169"  "s174"  "s177"  "s178"
[41] "s179"  "s181"  "s191"

> n.topID
[1] 41
> n.lowID
[1] 43

```

提示: `sort`, `which`

- (e) 試計算高分組及低分組在每一題答對的人數百分比, 記為 P_H 及 P_L 。

```

> PH
[1] 0.66 0.66 0.63 0.68 0.80 0.80 0.90 0.71 0.73 0.73
> PL
[1] 0.33 0.23 0.40 0.19 0.21 0.26 0.28 0.12 0.19 0.33

```

提示: `round`

- (f) 請計算每一題之難度 (公式 $P = (P_H + P_L)/2$) 及鑑別度 (公式 $D = P_H - P_L$)。

```

> P
[1] 0.50 0.44 0.52 0.44 0.50 0.53 0.59 0.42 0.46 0.53
> D
[1] 0.33 0.43 0.23 0.49 0.59 0.54 0.62 0.59 0.54 0.40

```

2.34 有一班學生之座號 (ID) 及性別 (`student.gender`) 的資訊如下。某日小考兩科: 微積分 (`score.calculus`) 及英文 (`score.english`)。成績如下, 其中有三位同學缺考。

```

set.seed(12345)
ID <- paste("No.", 1:50, sep="")
score.calculus <- sample(0:100, 50, replace=T)
score.english <- sample(0:100, 50, replace=T)
student.gender <- as.factor(sample(c("f", "m"), 50, replace=T))
absence.id <- sample(1:50, 3)
score.calculus[absence.id] <- score.english[absence.id] <- NA

```

- (a) 算出微積分平均分數及標準差。(提示: (1) 缺考不計入; (2) ?mean)
- (b) 男生英文成績平均多少分? (提示: 缺考不計入)
- (c) 將缺考成績記為 0 分後, 請問有哪些同學兩科成績同時及格? (列出座號)
- (d) (承上小題) 兩變數 $(x, y)_{i=1}^n$ 的相關係數之公式如下:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

試計算微積分及英文兩成績之相關係數, 並與 cor 之結果相比較。(提示: sqrt, sum, mean)

- 2.35 某校欲將學生之成績分組, 規則如下: 高於平均分數一倍標準差為「A」組, 低於平均分數一倍標準差為「C」組, 其餘為「B」組, 請將以下 30 位學生成績(score)依此規則分組。

```

set.seed(12345)
score <- sample(0:100, 30, replace=T)

```

- 2.36 將下列年齡資料(age)轉換為年齡群組(group), 規則如下: 1~20 歲為 A 組, 21~40 歲為 B 組, 41~60 歲為 C 組, 61 歲以上為 D 組。將轉換結果以資料框(data.frame)儲存, 使其第一個欄位為 age, 第二個欄位為 group。

```

set.seed(12345)
age <- sample(1:100, 20)

```

- 2.37 小吳老師於某系教授 A, B 兩班學生微積分, 學期各次成績使用同一格式紀錄於 score-A.txt 及 score-B.txt 兩檔案。檔案中紀錄 4 次小考成績、期中期末成績、助教(TA)成績, 各次考試之配分比例及學期點名出席次數。

- (a) 讀入兩資料檔, 將之合併為一個 data.frame (命名為 score), 使得各欄位名稱如下所示並增加一欄位註明班別(Class)。

	Class	No	ID	Name	Gender	Quiz1	Quiz2	Quiz3	Quiz4	TA	Midterm	Final	ATT
38	A	38	404550431	沈泓霏	女	15	25	53	67	93.3	29	42	9
39	A	39	404550442	許安霏	女	53	60	80	72	100.0	61	62	9
40	A	40	404550453	李政宜	男	80	100	85	100	100.0	95	100	3
41	B	1	404550465	史文羽	男	60	81	100	97	100.0	90	83	6
42	B	2	404685071	鄭樺妤	男	80	100	100	92	100.0	92	97	2
43	B	3	404685084	張敬安	男	10	40	62	93	100.0	65	84	9

- (b) 依各項考試 (小考、期中期末) 配分算出每位同學之學期成績 (缺考以零分計)。其中「出席成績」為額外加分，出席幾次，則總分加幾分。總分以不超過 100 為原則。請列出全班學期成績。
- (c) 列出學期成績在 55(含)~60 分 (不含) 之間的所有同學之全部各欄位紀錄。
- (d) A、B 兩班總成績平均各為多少？男、女生學期成績平均各為多少？
- (e) A 班學期成績不及格比例為多少？B 班男同學學期成績不及格比例為多少？
- (f) 分別印出男、女生學期成績前 5 名之「班別、學號、姓名、學期成績、名次」等欄位紀錄。(男、女生各按照名次依序列出) (名次為全班名次: `rank(x, ties.method = "first")`)
- (g) 印出「張」姓同學之完整姓名、學號及其學期成績。

2.38 美國大學成績平均績點 (GPA)(四分制) 的計算方式如下表:

等級 (Grade)	百分數	GPA
A	80 – 100 分	4
B	70 – 79 分	3
C	60 – 69 分	2
D	50 – 59 分	1
E	49 分以下	0

請寫一 R 函式，將某同學之各科修課成績百分數 (score) 轉成等級及 GPA。(提示: 不可用 `for`)

```
> set.seed(12345)
> score <- sample(0:100, 10, replace=T)
```

2.39 由螢幕輸入以下 10 個西元年份並由螢幕列印出來:

1224, 2065, 2000, 1660, 1020, 1986, 1787, 2080, 1147, 917

- (a) 印出最大及最小年份。
- (b) 小於 1500 的年份有哪些？

(c) 呈 (c) 小題，其平均年份及變異數為何？

2.40 利用 `for` 計算 `number` 中偶數的個數。

```
set.seed(12345)
number <- sample(0:100, 60, replace=T)
```

2.41 有一數學函數為

$$f(x) = \begin{cases} |x^2 + x|, & x < 0, \\ \sin(x), & 0 \leq x < 3, \\ 3e^x, & x \geq 3. \end{cases}$$

請寫一 R 函式，計算並列出下列表格：

	x	fx
1	-5	20.0000000
2	-4	12.0000000
3	-3	6.0000000
4	-2	2.0000000
5	-1	0.0000000
6	0	0.0000000
7	1	0.8414710
8	2	0.9092974
9	3	60.2566108
10	4	163.7944501
11	5	445.2394773

2.42 輸入包含左右小括號之字串 (最長為 40 字元)，請判斷是否左右小括號配對正確。

(例 1) 輸入： $((1+2)-3)*(4/5)$

輸出：括號配對正確。

(例 3) 輸入： $((1+2+3)$

輸出：括號配對不正確。

(例 3) 輸入： $((1+2)*(3+4)*(5+6))/(7+8)$

輸出：括號配對正確。

2.43 某國發行了 1 · 5 · 10 · 50 · 100 不同面額的鈔票，若有人要從銀行領出 N 元，銀行員要如何發給鈔票，使用的張數會最少？(試寫一 R 函式，命名為 `Change_Money`)

(例) 輸入: 478

輸出: 1 元 3 張, 5 元 1 張, 10 元 2 張, 50 元 1 張, 100 元 4 張, 共 478 元。

2.44 平面上兩點 $(x_1, y_1), (x_2, y_2)$ 之的距離式為: $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ 。給定 n 個點 ($n \leq 10$)，找出構成最小周長的三角形的三個點。

(例) 輸入: (1,1)(0,0)(4,3)(2,0)(7,8)

輸出: 三點為 (1,1)(0,0)(2,0)，其周長為 4.828428。

2.45 任意輸入 3 個座標，判別它是屬於下列哪種三角形: (1) 不是三角形 (2) 直角三角形 (3) 正三角形 (4) 等腰三角形 (5) 其它三角形。

例如: 三個座標為: (0, 0)(3, 0)(0, 4)

輸入: 0 0 3 0 0 4

輸出: 直角三角形

2.46 寫一 R 函式 (命名 `check_triangle`)，輸入為任意 3 個座標點，輸出為座標點所形成的三角形及三邊的長度，其中所形成的三角形有下列可能: (1) 不可為三角形，(2) 直角三角形，(3) 正三角形，(4) 等腰三角形，(5) 其它三角形。例如: 三個座標為: (0, 0)(3, 0)(0, 4)，則輸入: 「0 0 3 0 0 4」，並輸出: 「邊長: 3 4 5，直角三角形」。請利用以下五組座標點測試:

1. A: (4, 6) (-2, 8) (-8, 10)
2. B: (16, 14) (8, 2) (2, 6)
3. C: (0, 4) (8, 4) (4, 4+4\$\sqrt{3}\$)
4. D: (-4, 2) (4, 6) (-2, 8)
5. E: (12, 9) (8, 2) (2, 1)

2.47 試寫一 R 程式，由螢幕輸入三個座標點，判別這個三點是否可形成一三角形，若可以，則是屬於哪一種三角形 (純角、直角、銳角)。程式要求如下:

- (a) 需有: 標頭、使用者提示、輸出判別結果、是否繼續判別下一組座標點。(請參照程式風格講義範例 1)
- (b) 4 組測試座標: (1) (1,5), (3,1), (9,4); (2) (5,4), (2,1), (8,-3); (3) (3,4), (2,1), (1,-2); (4) (3,4), (2,1), (6,6).

2.48 世界衛生組織計算標準體重之方法如下:

$$\text{男性 : (身高 cm} - 80) \times 70\% = \text{標準體重}$$

女性 : (身高 cm - 70) × 60% = 標準體重

寫一 R 函式，命名為 ComputeWeight，沒有輸入。執行此程式後，會由螢幕詢問「性別」，及「身高 (公分)」，計算並印出此身高的標準體重 (公斤)。(執行 ComputeWeight 程式後，以 (1) 男生 175 公分及 (2) 女生 166 公分為範例)

- 2.49 某地購買物品需加收增值稅 (VAT)，而增值稅會根據所物品類別不同而有不同之稅率，今稅率表如下：

類別	產品	VAT
印刷品類	書, 雜誌, 報紙等等	8%
食物類	蔬菜, 肉品, 飲料等等	10%
衣服類	T 恤, 牛仔褲, 上衣等等	20%

某人買了下列 5 樣物品 (括弧內數字為不含稅之花費金額): {書 (50 元)、肉品 (200 元)、上衣 (299 元)、牛仔褲 (1200 元)、飲料 (20 元)}。(提示: %in%, ifelse)

- (a) 造出以下資料框 (data.frame)。

	品項	價格	類別	VAT
1	書	50	印刷品類	0.08
2	肉品	200	食物類	0.10
3	上衣	299	衣服類	0.20
4	牛仔褲	1200	衣服類	0.20
5	飲料	20	食物類	0.10

- (b) 請計算此次消費所付出之總稅額。

- 2.50 某醫生收集癌症病人的就診資料，其變數說明如下：

變數	名稱	說明
F1	編號	
F2	性別	1= 男 2= 女
F3	年齡	單位：歲
F4	發病天數	單位：月
F5	切片	1= 有 2= 沒有
F6	手術前的治療方式	0= 沒有 1= 化療 3= 化療加放療
F7	期別	
F8	手術日期	
F9	手術類型	
F10	追蹤日期	
F11	目前狀態	1= 沒有復發 2= 復發 5= 死亡 6= 不詳

- (a) 載入 `example.RData`，該工作空間中的 `cancer` 即為此筆資料集。請新增一資料集 `cancer0` 為 `cancer` 的複本，`cancer0` 為原始資料，`cancer` 為工作用資料，可拿來修改。
- (b) 利用指令 `for` 迴圈和 `print`，計算 `cancer` 資料中的類別變數 (F2, F5-F7, 及 F11) 各類別之個數。
- (c) 利用指令 `table`，計算 `cancer` 資料中類別變數 (F2, F5-F7, 及 F11) 各類別之個數。
- (d) 在 `cancer` 中，利用指令 `factor`, `levels` 將 F2, F5 中的 1,2 換成文字說明。利用指令 `ifelse` 將 F6, 及 F11 換成文字說明。計算類別變數 (F2, F5-F7, F11) 各類別之個數。(提示: `table`)
- (e) R 程式碼「`as.Date("6/20/94", format="%m/%d/%y")`」是將 1994 年 6 月 20 日的文字 (`character`) 類別轉成日期 (`Date`) 類別 "1994-06-20" (可用 `class` 去判斷)。日期記錄成 `Date` 類別之後，便可利用相減來得到天數。請將 F8 和 F10 轉成 `Date` 類別之後，再相減 (F10-F8) 得到追蹤天數 (命名為 F12)。印出 `cancer` 前 10 筆紀錄。
- (f) 在 `cancer` 資料集中，F12 或 F9 欄位有包含 `NA` 字樣，表示為遺失值。可利用 `is.na` 去判斷向量中的元素是否為遺失值。請把此資料中 F12 為 `NA` 的紀錄刪除。

2.51 有一 n 筆紀錄之資料標記為 $\{x_i, y_i\}_{i=1}^n$ ，其中 x_i 為一數值型變數觀察值， y_i 為一類別變數觀察值 (共 K 類別)，今欲計算此資料之 Fisher discriminant index (費雪區別指標， BSS/WSS)，其公式如下：

$$\frac{BSS}{WSS} = \frac{\sum_{i=1}^n \sum_{k=1}^K I(y_i = k)(\bar{x}_k - \bar{x})^2}{\sum_{i=1}^n \sum_{k=1}^K I(y_i = k)(x_i - \bar{x}_k)^2},$$

其中 $I(\cdot)$ 為指示函數 (indicator function)， \bar{x} 為 $\{x_i\}_{i=1}^n$ 之平均數， \bar{x}_k 為第 k 類別中， $\{x_i\}_{i=1}^n$ 之平均數。請寫一 R 函數，計算費雪區別指標，並以下列資料 (`x`, `y`) 為例。

```
set.seed(123456)
id <- sample(1:150, 30)
x <- iris[id, 1]
y <- iris[id, 5]
```

2.52 某樂透 (Lottery) 遊戲規則如下：「消費者從 01~49 中任選 6 個號碼進行投注。開獎時，開獎單位將隨機開出 6 個號碼 (winning number)。如果消費者選號有三個以上 (含三個號碼) 對中當期開出之 6 個號碼，即為中獎，並可依規定兌領獎金。」某天小明買了兩注電腦選號，其號碼為 (5, 29, 12, 10, 38, 35) 和 (41, 13, 21, 29, 19, 12)，若當期之開獎號碼為 (10, 7, 12, 38, 47, 35)，請寫一 R 函式，幫小明對獎。程

式要求如下: (1) 輸入為開獎號碼 (預設值為本題之開獎號碼); (2) 執行對獎程式後，由營幕輸入「消費者投注號碼」；(3) 輸出為消費者投注號碼及開獎號碼、對中之號碼個數、恭喜中獎或銘謝惠顧；(3) 不可用 `for`。(提示: (1) %*%; (2) 由營幕輸入「消費者投注號碼」，可一次輸入兩注，或一次輸入一注但執行兩次對獎程式)

2.53 樂透彩對獎程式：在 1~42 的整數中，樂透彩會開出 6 個號碼以及一個特別號，中獎規則以及獎額如下：

獎項	規則	獎金
頭獎	6 個號碼全中	1,000,000
二獎	6 個號碼中 5 個, 另一個中特別號	100,000
三獎	6 個號碼中 5 個	10,000
四獎	6 個號碼中 4 個	1,000
五獎	6 個號碼中 3 個	100

註：中頭獎的不能再被視為中三獎，餘類推。

- (a) 若當期開出之號碼為 38, 28, 18, 8, 5, 10。而特別號是 42。小銘買了一張彩卷，選號為 15, 1, 8, 18, 28, 38。請問有對中之號碼為何？對中號碼個數為幾個？
- (b) 小吳也買了樂透彩，所選 5 組號碼記錄在 (`mylist.txt`) 檔案。請你寫一 R 程式 `lotto` 幫他對獎，使得輸出為以下所列。(提示 1: 輸入為當期開出之號碼、特別號及號碼記錄檔。)

no.	中獎	累積獎金	有對到之號碼
=====			
1	沒中獎	0	(38)
2	中二獎	100000	(8 18 28 38 5 [42])
3	中五獎	100100	(8 18 28)
4	沒中獎	100100	()
5	中頭獎	1100100	(5 10 8 18 28 38)
=====			
總獎金：1100100 元			

(提示 2: `as.matrix`, `as.integer`, `if`, `for`, `cat`, `length`, `which`, ...)

2.54 USArrests 資料中，選出以 "N" 開頭的州，計算選出資料每個變數的平均值及標準差。

2.55 某公路經過 A, B, \dots, G 七個城市，各城市離出口之里程數依序為 25, 49, 95, 178, 264, 327, 373(公里)。現在要訂公車票價，規則如下

公里數	收費
50 公里內 (含)	一律收 100 元
50 公里以上 (不含) 且 在 300 公里 (含) 以內者	基本費 100 元加上 超過 50 公里的部份為每公里加收 1 元
超過 300 公里	一律收 400 元

請寫一 R 函式，輸入為城市離出口之里程數，輸出為城市間的票價表。(提示: matrix, for, if)

票價表	A	B	C	...	G
A	100	100	.	.	.
B	100	100	.	.	.
C	.	.	100	.	.
:
G	100

- 2.56 (a) 資料壓縮: 將字串 "AAABBBCCCC" 表示成 "3A3B4C"。 (提示: gregexpr, cat。)
 (b) 資料解壓縮: 將字串 "3A3B4C" 表示 "AAABBBCCCC"。 (提示: substr, cat, rep)

- 2.57 寫一「簡單資料壓縮」之 R 函式 (命名為 compress): 輸入為 ABC 三個字母組成之字串，例如：字串 "ABAABBAABCCAC"，輸出為每個字母出現之次數，(例如："6A4B4C")。(提示: gregexpr, for, LETTERS, paste0, length, cat。)

- 2.58 R 內建資料集 mtcars 是一汽車趨勢道路測試資料 (Motor Trend Car Road Tests) (1974 年)，資料包括 32 款汽車在油耗及 10 個汽車設計和性能測試相關的數據。11 個變數依序為：mpg (Miles/(US) gallon, 公哩/加侖), cyl (Number of cylinders, 氣缸數), disp (Displacement, 容量), hp (Gross horsepower, 總馬力), drat (Rear axle ratio, 後輪軸比) wt (Weight, 重量), qsec (1/4 mile time, ¼ 哩的時間), vs (Engine, 發動機類型), am (Transmission, 變速器) gear (Number of forward gears, 前進檔位數), carb (Number of carburetors, 化油器數)。請計算五個變數：disp, hp, drat, wt, qsec 在各個不同氣缸數之下的平均數。(提示：限使用一次 apply 及 tapply)。

- 2.59 身分證字號驗証規則如下：字母 (ABCDEFGHIJKLMNPQRSTUVWXYZ) 對應一組數 (10~35)。

縣市別	英文代號	數字編碼	縣市別	英文代號	數字編碼	縣市別	英文代號	數字編碼
台北市	A	10	新竹縣	J	18	高雄縣	S	26
台中市	B	11	苗栗縣	K	19	屏東縣	T	27
基隆市	C	12	台中縣	L	20	花蓮縣	U	28
臺南市	D	13	南投縣	M	21	台東縣	V	29
高雄市	E	14	彰化縣	N	22	澎湖縣	X	30
台北縣	F	15	雲林縣	P	23	陽明山	Y	31
宜蘭縣	G	16	嘉義縣	Q	24	嘉義市	I	34
桃園縣	H	17	台南縣	R	25	新竹市	O	35
			金門縣	W	32	連江縣	Z	33

令其十位數為 X_1 ，個位數為 X_2 ；(例如 A 的 $X_1 = 1$, $X_2 = 0$)，令 $D_1 \sim D_9$ 表示第 2~第 9 個數字，再令 $Y = X_1 + 9X_2 + 8D_1 + 7D_2 + 6D_3 + 5D_4 + 4D_5 + 3D_6 + 2D_7 + 1D_8 + D_9$ 。如 Y 能被 10 整除，則表示該身分證號碼為正確，否則為錯誤。請寫一身分證字號檢查的 R 程式 (命名為 check_id)，輸入為檔名 (id.txt 紀錄 5 筆台灣身分證字號)，輸出為以下表格。

身份字號	數字編碼	縣市別	Y 值	正確性 (Y/N)
<hr/>				
F183741875				
A148992712				
T189179230				
P139392302				
H146359668				

2.60 小銘雞排國際股份有限公司三峽分部提供給員工使用的資料庫存取訊息如下：
MySQL Server IP: 163.13.113.xxx, port=3306; 資料庫名稱: bigdata105; 使用者帳號: student; 密碼: xxxxxxxx。

- (a) 請將資料表格”student_info”讀入 R 後，依照學號排序 (遞增)，刪除重覆之紀錄，列出資料前六筆紀錄。(刪除後) 共有多少筆紀錄？共有多少欄位？
- (b) BMI(身體質量指數) 值計算公式為¹: $BMI = \frac{\text{體重}}{\text{身高}^2}$ ，其中體重單位是公斤，身高單位是公尺。請由螢幕輸出提示給使用者「請輸入您的體重 (公斤)/身高 (公尺)」，由鍵盤輸入「體重 (公斤)/身高 (公尺)」，計算 BMI 後，再由螢幕輸出 BMI 值「您的 BMI 值為: 」。
- (c) BMI(身體質量指數) 值計算公式為²: $BMI = \frac{\text{體重}}{\text{身高}^2}$ ，其中體重單位是公斤，身高單位是公尺。依所計算出的 BMI，將體重判別分類如下：

若 $BMI < 18.5$ ，則表示「體重過輕」；
 若 BMI 介於 18.5(含) 和 24 之間，則表示「體重正常」；
 若 BMI 介於 24(含) 和 27 之間，則表示「體重過重」；
 若 BMI 介於 27(含) 和 30 之間，則表示「輕度肥胖」；
 若 BMI 介於 30(含) 和 35 之間，則表示「中度肥胖」；
 若 $BMI > 35$ (含)，則表示「重度肥胖」；

¹<http://depart.femh.org.tw/dietary/3OPD/BMI.htm>

²<http://depart.femh.org.tw/dietary/3OPD/BMI.htm>

試寫一 R 函式，輸入為「身高及體重」，輸出為「BMI 值及體重判別」。並以上小題之資料為例，印出每個人之姓名、體重、身高、BMI 值，及其體重判別。

2.61 資料檔 `wine.csv` 紀錄三種不同酒類 (Type) 的 13 種化學成份指數 (變數)，共有 178 個酒的樣本 (編號依序為 1~178)。

- (a) 讀取資料檔，印出資料前 5 筆及後 5 筆紀錄。
- (b) 計算每一種酒類的 13 個化學成份指數之平均數。
- (c) 印出每一化學成份指數最高之酒的編號。

2.62 某一公司 2018 年之銷售紀錄 (saleRecord) 檔如下，包含三個欄位 (銷售日期、銷售物品、銷售量)：

```
n <- 100
date.period <- seq(as.Date('2018-1-1'),
                      to = as.Date('2018-12-31'), by = '1 day')
x <- sample(date.period, n, replace = T)
y <- sample(LETTERS[1:4], n, replace = T)
z <- floor(rnorm(n, m=100, sd=10))
saleRecord <- data.frame(saleDate=x, saleItem=y, saleVolume=z)
```

計算 2018 年各季 (第一季到第四季) 之各物品 (A, B, C, D) 的銷售總次 (日) 數及銷售量總和。

2.63 試寫一 R 程式，實作 k 最近鄰居分類法 (k -Nearest Neighbor Classifier, KNN)。

- 輸入：
 - `x.train`: 維度為 $n \times p$ ，是一具有 n 個觀察值 p 個變數的訓練集資料矩陣 (x_1, x_2, \dots, x_n) 。
 - `y.train`: 長度為 n ，是訓練集資料每個觀察值的類別 (y_1, y_2, \dots, y_n) ，具有 g 個類別。
 - `k`: 最近鄰居個數。預設值為 5。
 - `x.test`: 維度為 $m \times p$ ，是一具有 m 個觀察值 p 個變數的測試集資料矩陣 $(x_1^t, x_2^t, \dots, x_m^t)$ 。
- 輸出: `y.pred`: 長度為 m ，是以 KNN 分類測試集資料矩陣所得到的預測類別。
- 演算法:
 - 計算 `x.test` 中第 1 個觀察值 (x_1^t) 到 `x.train` 中每一個觀察值的距離 (d_1, d_2, \dots, d_n) 。

- 於上述 n 個距離中，選出距離最近的 $k = 5$ 個觀察值 $(x^{(1)}, x^{(2)}, \dots, x^{(5)})$ 。
- 上述 $k = 5$ 個觀察值，其相對應的類別為 $(y^{(1)}, y^{(2)}, \dots, y^{(5)})$ 。
- 判別 x_1^t 的類別為上述各類別總數最多者。
- 以 `x.test` 中第 2 個觀察值 (x_2^t) 重覆第一步驟，直到 `x.test` 裡所有觀察值皆判別完畢。
- 不要用 `for`。指令提示: `table`, `unique`, `which`, `sort`, `order`, `dist`.

```

set.seed(12345)
id <- sample(1:150, 100)
x.train <- iris[id, 1:4]
y.train <- iris[id, 5]
x.test <- iris[-id, 1:4]
myKNN <- function(...){
  ...
  ...
}

myKNN(...)

```

2.64 假設資料中兩個變數分別記做 X 跟 Y ，它們的元素個數均為 n ，其第 i 個值分別用 (x_i, y_i) 表示 ($1 \leq i \leq n$)。以下針對任意 (x_i, y_i) 與 (x_j, y_j) 定義三種狀況 ($1 \leq i, j \leq n$):

- (x_i, y_i) 與 (x_j, y_j) 為同一配對：滿足 $\{x_i > x_j \text{ 且 } y_i > y_j\}$ ，或 $\{x_i < x_j \text{ 且 } y_i < y_j\}$ 。
- (x_i, y_i) 與 (x_j, y_j) 為不同一配對：滿足 $\{x_i > x_j \text{ 且 } y_i < y_j\}$ ，或 $\{x_i < x_j \text{ 且 } y_i > y_j\}$ 。
- (x_i, y_i) 與 (x_j, y_j) 為非同一配對，也非不同一配對： $\{x_i = x_j\}$ ，或 $\{y_i = y_j\}$ 。

以 `x <- iris[,1]; y <- iris[,3]` 為例，寫一 R 函式 (命名為 `pairs_cal`)，計算並輸出此資料的「同一配對個數」與「不同一配對個數」。

2.65 建立一 R 套件專案，名為「學號-R-exam3-2」。

- 於此 R 套件中新增一 R 程式碼檔，名為「`source.R`」。於此程式碼檔案，實作下列程式：

假設資料中兩個變數分別記做 X 跟 Y ，它們的元素個數均為 n ，其第 i 個值分別用 (x_i, y_i) 表示 ($1 \leq i \leq n$)。以下針對任意 (x_i, y_i) 與 (x_j, y_j) 定義三種狀況 ($1 \leq i, j \leq n$):

- i. (x_i, y_i) 與 (x_j, y_j) 為同一配對 (concordant pairs): 滿足 $\{x_i > x_j \text{ 且 } y_i > y_j\}$ ，或， $\{x_i < x_j \text{ 且 } y_i < y_j\}$ 。
- ii. (x_i, y_i) 與 (x_j, y_j) 為不同一配對 (discordant pairs): 滿足 $\{x_i > x_j \text{ 且 } y_i < y_j\}$ ，或， $\{x_i < x_j \text{ 且 } y_i > y_j\}$ 。
- iii. (x_i, y_i) 與 (x_j, y_j) 為非同一配對，也非不同一配對：即 $\{x_i = x_j\}$ ，或， $\{y_i = y_j\}$ 。

Kendall 等級相關係數 (Kendall rank correlation coefficient) 的計算公式如下：

$$\tau = \frac{\{\text{同一配對個數}\} - \{\text{不同一配對個數}\}}{n(n-1)/2}.$$

以 `x <- iris[, 1]; y <- iris[, 3]` 為例，寫一 R 函式 (命名為 `kendall_tau`) 計算 Kendall 等級相關係數，並與 `cor.` 相比較。

- (b) 寫出此程式之說明文件，需至少包含 Description, Usage, Arguments, Details, Value, See Also, 和 Examples。
 - (c) 編譯此套件專案，並製出此套件之二元安裝檔 (學號-R-exam3-2_0.1.0.zip)。
- 2.66 寫一「剪刀石頭布遊戲」的 R 程式。執行畫面示意如下。(提示: (1) 你的答案可能跟畫面不一樣。(2) 電腦出拳是隨機抽樣。(3) 畫面至少玩 8 次以上，最後一次是「不玩了」)

剪刀石頭布遊戲開始

請輸入你要出的拳頭

(a: 剪刀, b: 石頭, c: 布, d: 不玩了): a
電腦出布 · 你出剪刀 · 你贏了!

請輸入你要出的拳頭

(a: 剪刀, b: 石頭, c: 布, d: 不玩了): b
電腦出石頭 · 你出石頭 · 你們平手!

請輸入你要出的拳頭

(a: 剪刀, b: 石頭, c: 布, d: 不玩了): c
電腦出剪刀 · 你出布 · 你輸了!

請輸入你要出的拳頭

(a: 剪刀, b: 石頭, c: 布, d: 不玩了): d
謝謝再會!

2.67 依下列步驟，完成一「剪刀石頭布遊戲」的 R 程式。

- (a) (10 分) 由電腦隨機產生一個拳頭 (剪刀、石頭、布)，並印出。(提示: `sample`)
 (b) (10 分) 由螢幕輸入玩家要出的拳頭，使得執行的畫面如下。(提示: `switch`)

請輸入你要出的拳頭(a: 剪刀, b: 石頭, c: 布, d: 不玩了): a
玩家出: 剪刀

- (c) (50 分) 寫一「剪刀石頭布遊戲」的 R 程式 (命名 `game`)，使得程式執行的畫面如下。(提示: (1) 電腦出拳是隨機抽樣。(2) `repeat`, `break`)

```
> set.seed(12345)
> game()
### 剪刀石頭布遊戲開始 ####
請輸入你要出的拳頭(a: 剪刀, b: 石頭, c: 布, d: 不玩了):
1: a
電腦出[ 布 ], 你出[ 剪刀 ], 你[ 贏 ]了!

請輸入你要出的拳頭(a: 剪刀, b: 石頭, c: 布, d: 不玩了):
1: b
電腦出[ 布 ], 你出[ 石頭 ], 你[ 輸 ]了!

請輸入你要出的拳頭(a: 剪刀, b: 石頭, c: 布, d: 不玩了):
1: c
電腦出[ 布 ], 你出[ 布 ], 你[ 平手 ]了!

請輸入你要出的拳頭(a: 剪刀, b: 石頭, c: 布, d: 不玩了):
1: d
謝謝再會!
```

2.68 小明和小漢在玩 5×5 的數字賓果遊戲。開獎數字報出後，賓果盤上相對應的數字則以加記星號表示，若某一橫列或直列或對角列之 5 個數字皆被標記，則記為一連線。誰先得到五連線則為贏家。程式設計要點如下：

- 請隨機產生兩個並排之數字賓果盤（數字 1 至 25 擺至 5×5 之矩陣不重覆，你的答案和以下所列可能不同）。
- 請隨機產生一個開獎數字，兩個人之賓果盤上相對應的數字則以加記星號表示。
- 重覆上述開獎過程，開獎數字與之前已開出之號碼不重覆。
- 計算連線數，若達到設定連線數，則為贏家。

設定本數字賓果遊戲先達成之連線數為贏家: 1

小明 小漢

=====

7	14	16	9	24	13	4	7	20	9
4	6	22	17	1	5	11	15	6	17
18	12	19	25	11	12	14	24	3	25
8	15	20	21	13	22	2	21	19	8
5	2	3	23	10	10	23	18	16	1

繼續開獎(y/n): y 開獎號碼: 4

小明 小漢

=====

7	14	16	9	24	13	4*	7	20	9
4*	6	22	17	1	5	11	15	6	17
18	12	19	25	11	12	14	24	3	25
8	15	20	21	13	22	2	21	19	8
5	2	3	23	10	10	23	18	16	1

繼續開獎(y/n): y 開獎號碼: 13

小明 小漢

=====

7	14	16	9	24	13*	4*	7	20	9
4*	6	22	17	1	5	11	15	6	17
18	12	19	25	11	12	14	24	3	25
8	15	20	21	13*	22	2	21	19	8
5	2	3	23	10	10	23	18	16	1

....

繼續開獎(y/n): y 開獎號碼: 19

小明 小漢

=====

7	14	16	9	24*	13*	4*	7	20	9
4*	6	22	17	1	5	11*	15	6	17
18	12	19*	25	11*	12	14	24*	3	25
8	15	20	21	13*	22	2	21	19*	8
5	2	3	23	10	10	23	18	16	1*

小漢: 1 連線，小明: 0 連線。小漢 為贏家。遊戲結束。

2.69 小銘和小漢在玩「幾 A 幾 B 猜數字」的遊戲。若小銘真正答案為「2985」，小漢猜測為「1928」，即為「1A2B」，請幫小銘寫一 R 程式自動報出幾 A 幾 B，直到小漢

猜測全答對為止。(提示: 讀取猜測數字 \Rightarrow 判別 \Rightarrow 報出幾 A 幾 B \Rightarrow 若為 4A 則程式結束, 否則再次讀取猜測數字(迴圈))

幾A 幾B猜數字: 小銘答案: 2985

```
=====
小漢猜測: 1928 => 1A2B
小漢猜測: 2934 => 2A
小漢猜測: 2958 => 2A2B
小漢猜測: 2985 => 4A
=====
```

- 2.70 小魯哥在玩「幾 A 幾 B 猜數字」的遊戲, 規則是電腦隨機產生一組 4 位數, 其數字為 1~9, 不得重覆。(本題設定是 1~9 喔!) 若假設產生之數字為「8739」, 小魯哥猜測為「1938」, 即為「1A2B」, A 代表某一數字及位置正確, B 代表某一數字正確但位置不正確, AB 字母前面的數字代表這種類型的數字有幾個; X 代表 4 個數字皆不正確。(若不清楚遊戲規則, 可自行 google 一下。) 請幫小魯哥寫一 R 程式(命名為 play, 輸入為學號、姓名及隨機種子)自動報出幾 A 幾 B, 直到小魯哥猜測全答對為止。(註: 需測試兩組以上的隨機種子; 需印出是第幾次猜測。)

```
play <- function(id, name, seed){
  set.seed(seed)
  computer <- ...
  ... scan ....
}

> play(410971234, "小魯哥", 123456)
學號: 410971234; 姓名: 小魯哥
幾A幾B猜數字遊戲:
=====
第1次猜測: 1245 => X
第2次猜測: 1938 => 1A2B
第3次猜測: 2938 => 1A2B
第4次猜測: 1839 => 2A1B
第5次猜測: 8739 => 4A (遊戲結束)
=====

> play(410971234, "小魯哥", 654321)
...
```

2.71 小銘到巷口跟賣香腸的阿伯玩十八啦，亦即擲四顆公平的六面骰子到一個大湯碗中，計算點數和，跟阿伯比大小，贏的話就可得到一根香腸。其中點數計算規則如下：

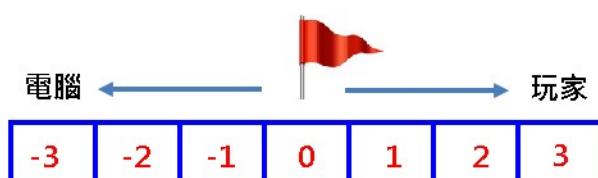
- 四個骰子挑出兩顆相同的不計，看另兩顆骰子點數和：例：3345，33 成對拿掉，剩下 45，點數和為 9。例：1662，66 成對拿掉，剩下 12，點數和為 3，直接判定最輸，這個叫「逼基」。
- 若四顆點數皆異（如 2456），或有三顆點數相同（如 1555），都不算，重新擲骰子。
- 若兩兩相同（如 3344），取大的對子，即 44，點數和為 8。
- 兩顆六點 + 另兩顆相同的點數，例：3366, 2266, 1166 等，點數和為 12 點，這個叫「十八」。
- 四顆點數皆同（如 5555），稱為「豹子」、「通殺」或「一色」，點數和是最大。

請寫一 R 程式，模擬擲四個骰子的狀況，亦即隨機產生四個 1 至 6 的數字，當成擲四個骰子的點數，依上述規則，印出點數和或俗稱。請印出擲 10 次之結果。

四顆骰子點數	點數和/俗稱
<hr/>	
3345	9
1662	逼基
...	...
...	...
5555	通殺
2654	無面
2266	十八

2.72 **搶旗遊戲**：遊戲一開始，有一個旗子位於中央位置（編號 0），如下圖所示。玩家出拳（剪刀、石頭、布），與電腦相比，用猜拳來決定前進或後退。若玩家贏則往右一步，反之若電腦贏則往左一步，平手的話，則不動。若前進到位置 3 或後退到位置 -3，則遊戲結束，印出勝利者。

（註：R 函式命名 playEX3；需先 set.seed(123456)）



搶旗畫面程式：

```

show <- function(k=4){

  odd <- seq(1, 15, 2)
  even <- seq(2, 15, 2)
  flag <- character(15)
  flag[1:15] <- " "
  flag[even] <- "   "
  flag[even[k]] <- " f"

  loc <- character(15)
  loc[odd] <- "|"
  loc[even] <- c(-3:(-1), "00", paste0("+", 1:3))

  cat(flag, "\n")
  cat(loc, "\n")

  k
}

```

執行畫面示意範列如下頁:

```

> set.seed(123456)
> playEx3()
搶旗遊戲
      f
| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

(1) 剪刀 (2)石頭 (3) 布:
1: 1
電腦: 布
玩家: 剪刀
      f
| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

```

(1) 剪刀 (2)石頭 (3) 布:

1: 1

電腦: 布

玩家: 剪刀

f

| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

(1) 剪刀 (2)石頭 (3) 布:

1: 1

電腦: 石頭

玩家: 剪刀

f

| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

(1) 剪刀 (2)石頭 (3) 布:

1: 2

電腦: 石頭

玩家: 石頭

(1) 剪刀 (2)石頭 (3) 布:

1: 3

電腦: 石頭

玩家: 布

f

| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

(1) 剪刀 (2)石頭 (3) 布:

1: 2

電腦: 剪刀

玩家: 石頭

f

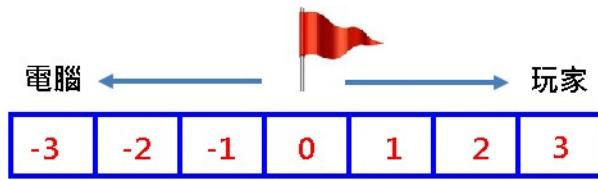
| -3 | -2 | -1 | 00 | +1 | +2 | +3 |

game over!

>

2.73 小魯哥考完電概 R 期中考後，覺得很沮喪，經過小胖老師的開導之後，小魯哥決定擺脫佛系學習法，開始奮發圖強。於是想到把期中考的「搶旗遊戲」，以 rpanel 實做出來。他的想法如下頁圖形所示，請你和他一起來完成這個專案吧!!

搶旗遊戲: 遊戲一開始，有一個旗子位於中央位置 (編號 0)。如下圖所示。玩家出拳 (剪刀、石頭、布)，與電腦相比，用猜拳來決定前進或後退。若玩家贏則往右一步，反之若電腦贏則往左一步，平手的話，則不動。若前進到位置 3 或後退到位置 -3，則遊戲結束，印出勝利者。



實作要求:

- (a) 建立一 R 專案，命名為「學號-R-HW5」。(注意: 專案不包含「姓名」)
- (b) 於專案內，新增一 R 程式檔，命名為「學號-姓名-R-HW5.R」。
- (c) 搶旗遊戲主程式 (R 函式)，命名為「play」。**(助教會執行測試!!)**
- (d) 於答案卷「學號-姓名-R-HW5.docx」貼上「rpanel 主控台」及數個「執行畫面」。並將答案卷置於專案內。(不需貼程式碼)
- (e) 如何擷取畫面:
 - i. 擷取「rpanel 主控台」：點選視窗，同時按「Alt + PrintScreen」，再將畫面用「Ctrl+V」貼到答案卷。
 - ii. 在 R 裡擷取「執行畫面」：檔案 => 複製到剪貼簿 => 做為 Bitmap；再將畫面用「Ctrl+V」貼到答案卷。
 - iii. 在 RStudio 裡擷取「執行畫面」：Export => Copy to Clipboard => Copy Plot；再將畫面用「Ctrl+V」貼到答案卷。
- (f) 將專案目錄「學號-R-HW5」壓縮成一壓縮檔「學號-R-HW5.zip」或「學號-R-HW5.rar」上傳。
- (g) 請你儘量達到小魯哥要求的功能，其它部份可自行發揮 (介面設計不同，或各元件不在主控台等等)，沒有標準答案。



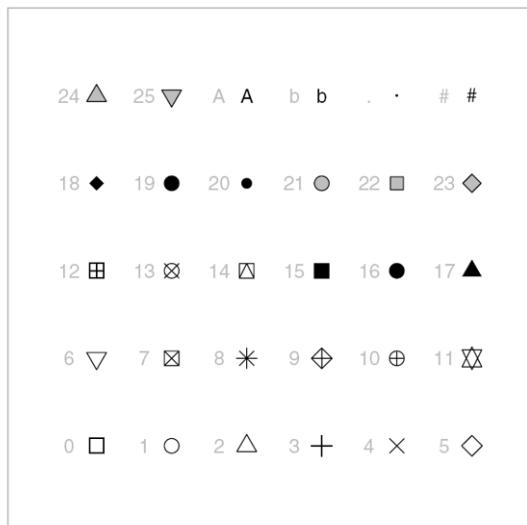
3 繪圖: Base graphics

3.1 畫出 $y = \log_a(x)$ $x > 0$ 之圖形。

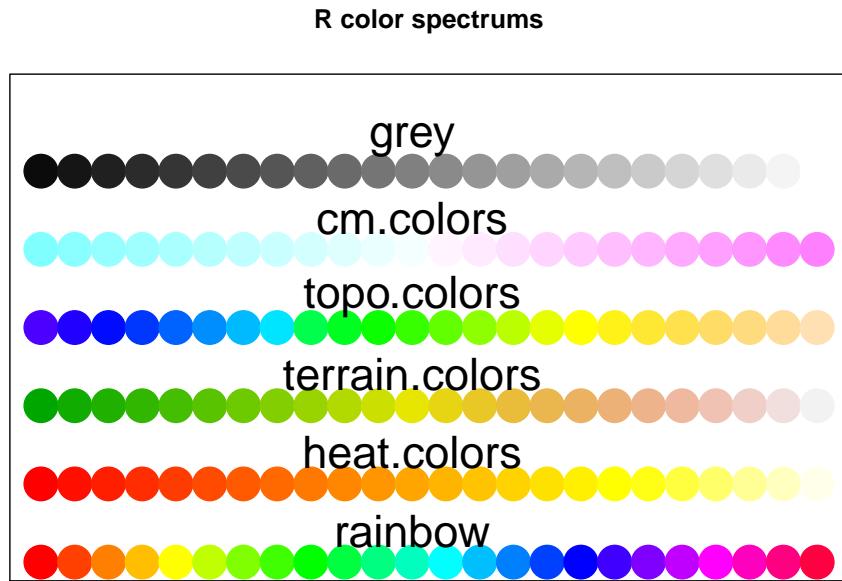
3.2 若有一函數為 $f(x) = \frac{\sin(kx)}{\sqrt{k}}$ ，畫出當 $k = 2, 4, 9$ 時之函數圖形 (需以不同顏色呈現)。

3.3 利用極座標 $x = 1.5 \cos(\theta) - \cos(30\theta)$, $y = 1.5 \sin(\theta) - \sin(30\theta)$, 繪出其圖形。

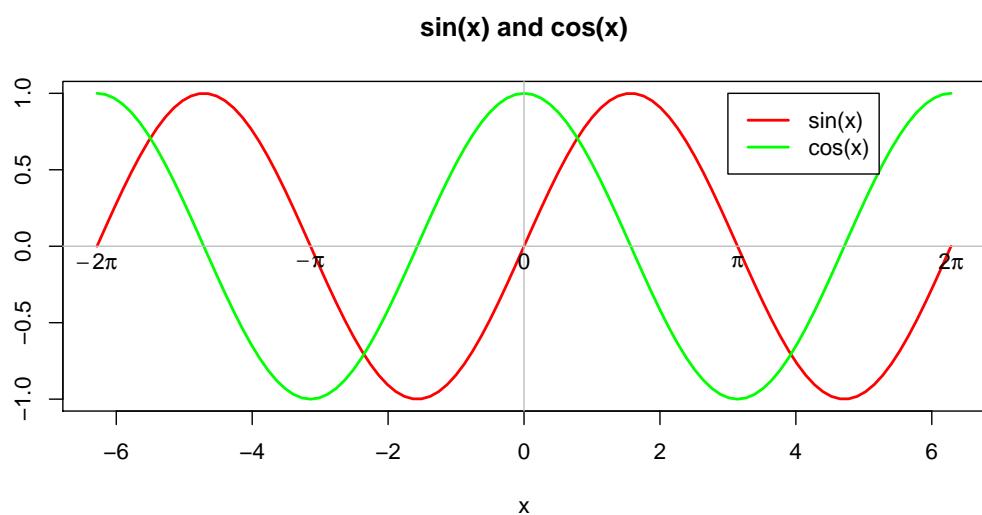
3.4 畫出下圖



3.5 畫出下圖。(限用一次 `plot` 及 `text` 指令)

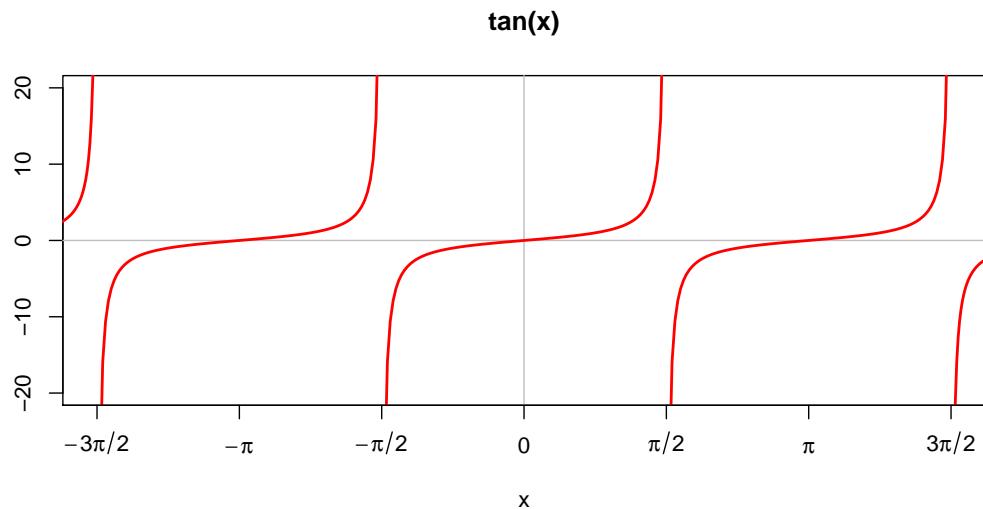


3.6 畫出 \sin 及 \cos 函數如下圖。(提示: `curve`, `abline`, `expression`, `text`)



3.7 畫出 \tan 函數如下圖。(提示 1: `plot`, `curve`, `axis`)

(提示 2: `points.at <- c(-2*pi, -3*pi/2, -pi/2, pi/2, 3*pi/2, 2*pi)`)



3.8 在標準常態分配的 density function 下，用紅色填滿小於 $z_{0.025}$ 和大於 $z_{0.975}$ 的區域。(提示: `dnorm`, `-1.96, 1.96`。)

3.9 $z \sim N(0, 1)$, 完成下圖。

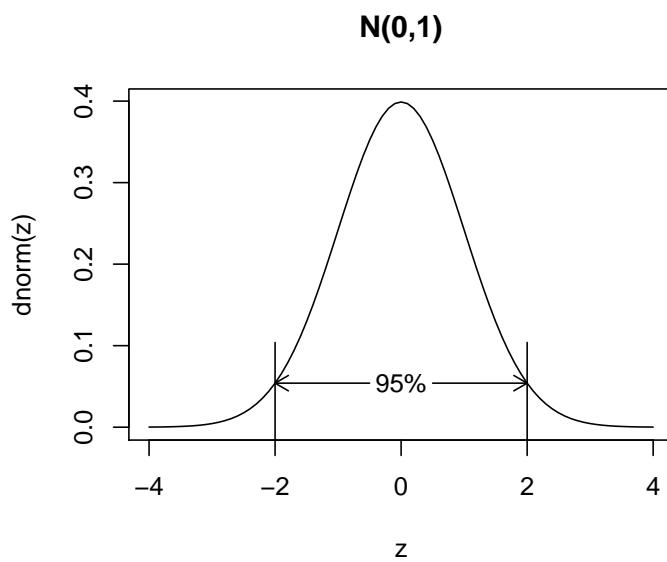
(a) 畫出標準常態分配的 density 圖。

(b) 加上 (紅色) 線段。

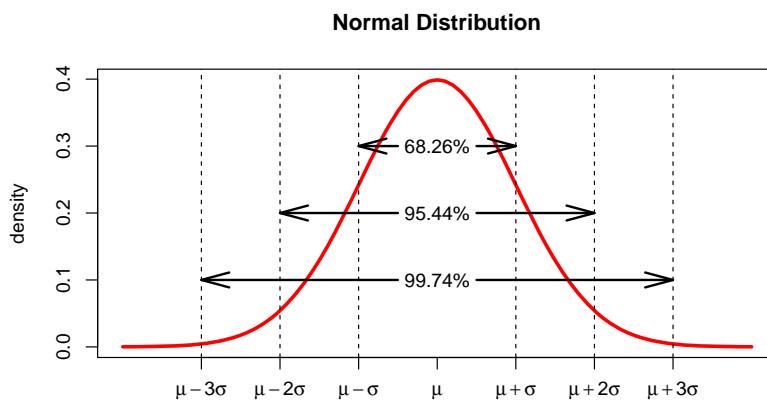
(Hint: `segments`。)

(c) 加上 (紅色) 標線及文字。

(Hint: `arrows`, `text`。)



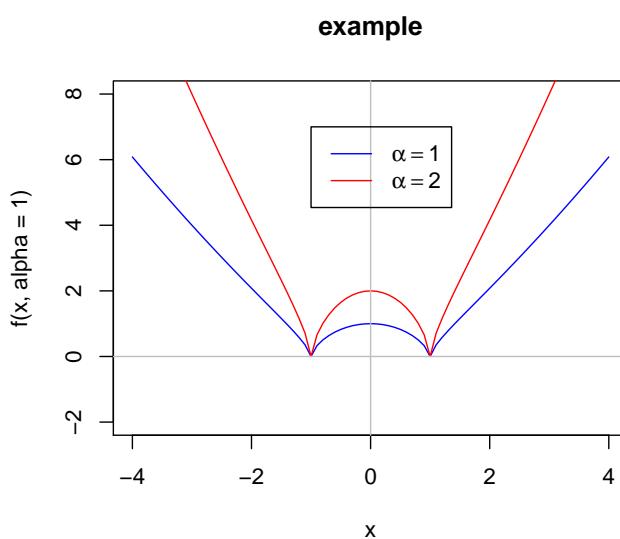
3.10 下圖紅色曲線為標準常態機率密度函數 (`dnorm`)，請畫出下圖。(限用兩次 `arrows`，一次 `text`，一次 `abline`，詳見提示。)



提示:

```
x <- ...
y <- ...
plot(x, y, ...
arrows(..., ...
arrows(..., ...
text(..., ...
abline(..., ...
axis(1, at=c(-3:3), labels=c(expression(mu-3*sigma), ...
```

3.11 畫出函數 $f(x) = \alpha(x^2 - 1)^{2/3}$ 圖形如下:

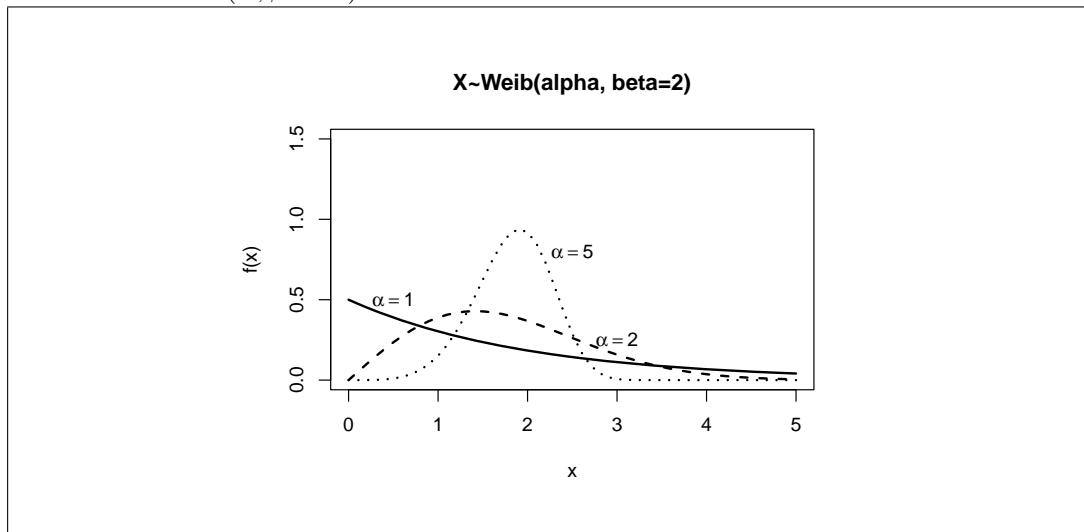


3.12 若隨機變數 X 服從 Weibull 分配 (簡記為 $X \sim Weib(\alpha, \beta)$) · 其機率密度函數為

$$f(x|\alpha, \beta) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha}, \quad x > 0.$$

(a) 若 $x \leftarrow \text{seq}(0, 5, 0.1)$ · 寫一函式計算 $f(x|\alpha = 1, \beta = 2)$ 之值。

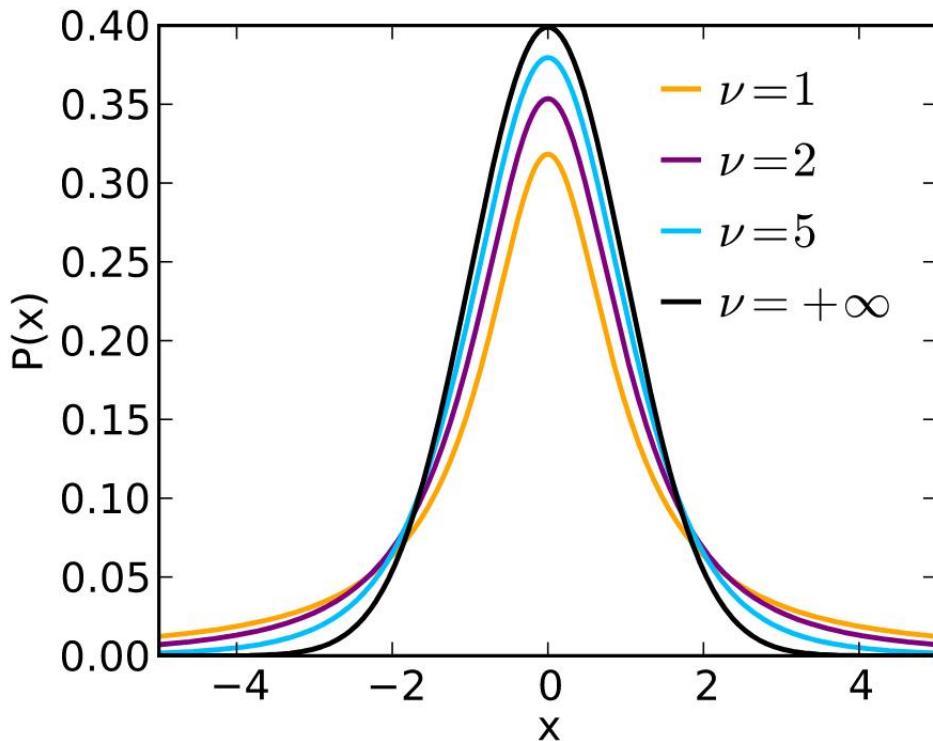
(b) 畫出 $X \sim Weib(\alpha, \beta = 2)$ 之圖形如下:



3.13 Student's t -distribution has the probability density function (pdf) given by

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function. 試寫一 R 函數 (命名為 `t_pdf` · 輸入為 x 、 ν · 輸出為 $p(x)$) · 計算 t 分配之機率密度函數值 · 並利用此函數畫出下圖 (注意: x -、 y - 軸標號、圖例說明 (legend)、線條顏色)。(註: Γ 函數在 R 是什麼指令 · 請自己查。 $\nu = +\infty$ 以 $\nu = 10$ 代入。畫圖以 R base graphics 套件或 ggplot2 皆可。) (圖片來源: https://en.wikipedia.org/wiki/Student%27s_t-distribution



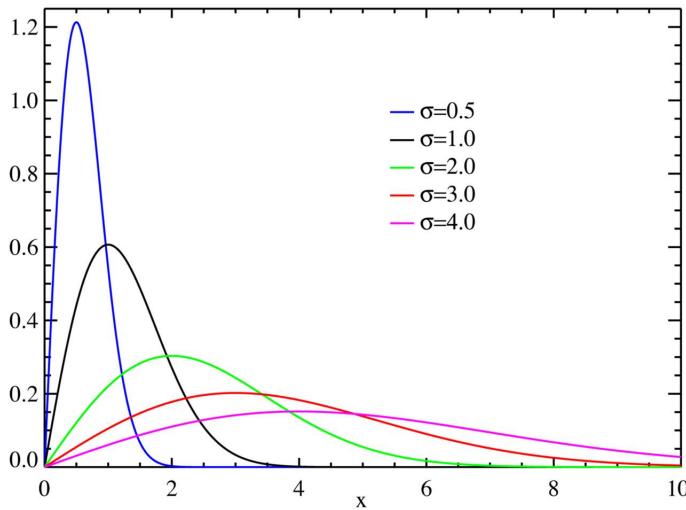
3.14 t 分佈在 wikipedia 中的介紹如下列網頁: https://en.wikipedia.org/wiki/Student%27s_t-distribution。以 R 基礎套件 (base graphics) · 畫出 t 分佈在自由度為 1 及 自由度為 5 的 (a) 機率密度函數圖 · (b) 累積機率分佈函數圖 · (c) 分位數函數圖及 (d) 隨機抽樣 ($n = 100$) 直方圖。(要求: (1) 前三個圖上各有兩條不同自由度之函數 曲線 (以不同顏色表示) · 直方圖則為重疊 (以不同顏色表示)。 (2) 需加註: 標題 · x 及 y 標號及 legend (以不同顏色表示相對應的自由度)。 (3) 4 張圖一頁: 2 by 2)。

3.15 The probability density function of the Rayleigh distribution is:

$$f(x; \sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x \geq 0,$$

where σ is the scale parameter of the distribution³. 畫出此分佈的 Probability density functions 圖形如下:

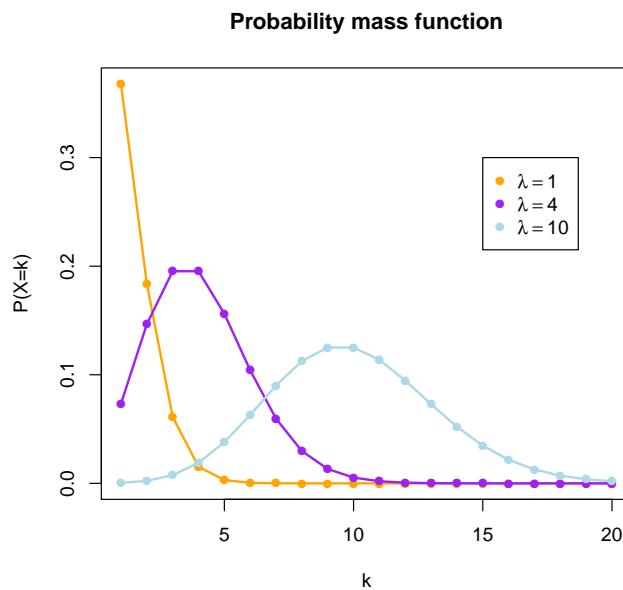
³https://en.wikipedia.org/wiki/Rayleigh_distribution



3.16 卜瓦松分布 (Poisson distribution) 的機率質量函數 (Probability mass function) 為

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

利用 `matplotlib` 畫出機率質量函數圖如下 (https://en.wikipedia.org/wiki/Poisson_distribution)。
(提示: `c("orange", "purple", "lightblue")`, `data.frame`, `type="o"`, `expression`)

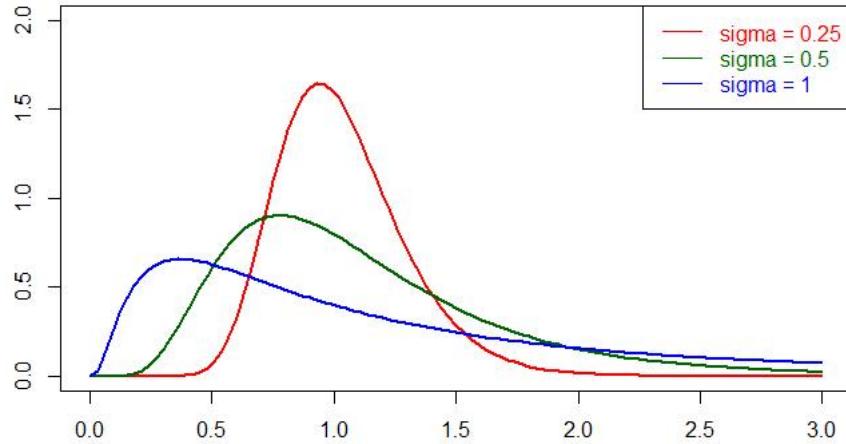


3.17 紿定 $\lambda = 2.4$ 及 $X_n \sim B(n, \lambda/n)$ ，分別畫出當 $n = 10, 30, 100$ 時， X_n 之機率質量函數圖，以及 $X \sim Poisson(2.4)$ 之機率質量函數圖。(提示: 一頁 4 張圖，各圖需加上適合之標題及 xy 軸標號)

3.18 有一函數 $f(x; \mu, \sigma)$, 其數學式為

$$f(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0, \mu \in R, \sigma > 0.$$

- (a) 寫一 R 函式實作此函數 (命名為 `my.pdf`)，其中輸入參數 μ 的預設值為 0, σ 之預設值為 1。
- (b) 依據上小題，繪出下圖。



3.19 伽瑪分布 (Gamma Distribution) 是統計學的一種連續機率函數 (https://en.wikipedia.org/wiki/Gamma_distribution)，

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}, \quad x > 0,$$

其中參數 α 稱為形狀參數 (shape), θ 稱為尺度參數 (scale), $\Gamma(\alpha)$ 為 Gamma 函數 (R 指令為 `gamma`)，滿足

$$\Gamma(\alpha) = \begin{cases} (\alpha - 1)! & \text{if } \alpha \in \mathbb{Z}^+, \\ (\alpha - 1)\Gamma(\alpha - 1) & \text{if } \alpha \in \mathbb{R}^+. \end{cases}$$

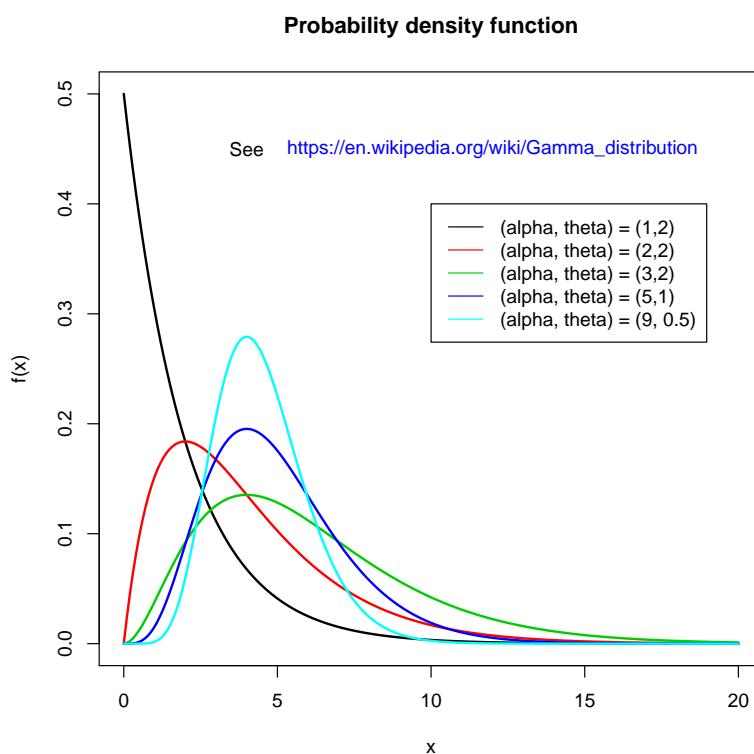
- (a) 以 R 程式實作伽瑪分布 (命名為 `my_dgamma`)，輸入為 `(x, alpha, theta)`，輸出為 $f(x)$ 值。
- (b) 若 α 及 θ 為下列各配對值 (即 $\alpha = 1, \theta = 2; \alpha = 2, \theta = 2; \dots; \alpha = 9, \theta = 0.5$)，且當 $x = 0, 0.1, 0.2, \dots, 20$ 時，計算機率密度函數 $f(x)$ 值 (儲存成一 R 矩陣物件，命名為 `dgamma.table`)，並列印出此表格前 10 筆資料如下。(指令提示: `matrix`, `for`, `alpha[i]`, `theta[i]`)

```

> x <- seq(0, 20, 0.1)
> alpha <- c(1, 2, 3, 5, 9)
> theta <- c(2, 2, 2, 1, 0.5)
...
...
> dgamma.table[1:10, ]
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.5000000 0.00000000 0.0000000000 0.000000e+00 0.000000e+00
[2,] 0.4756147 0.02378074 0.0005945184 3.770156e-06 1.039658e-10
[3,] 0.4524187 0.04524187 0.0022620935 5.458205e-05 2.179072e-08
[4,] 0.4303540 0.06455310 0.0048414824 2.500261e-04 4.572385e-07
[5,] 0.4093654 0.08187308 0.0081873075 7.150080e-04 3.739330e-06
[6,] 0.3894004 0.09735010 0.0121687622 1.579507e-03 1.824799e-05
[7,] 0.3704091 0.11112273 0.0166684100 2.963583e-03 6.424008e-05
[8,] 0.3523440 0.12332042 0.0215810727 4.967922e-03 1.805184e-04
[9,] 0.3351600 0.13406401 0.0268128018 7.668548e-03 4.301284e-04
[10,] 0.3188141 0.14346633 0.0322799252 1.111460e-02 9.035651e-04

```

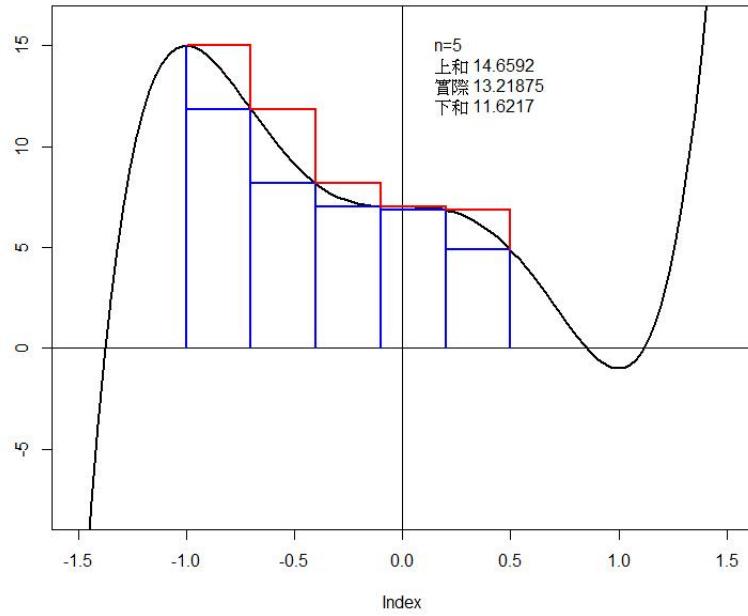
(c) 利用 `dgamma.table`，畫出下圖。



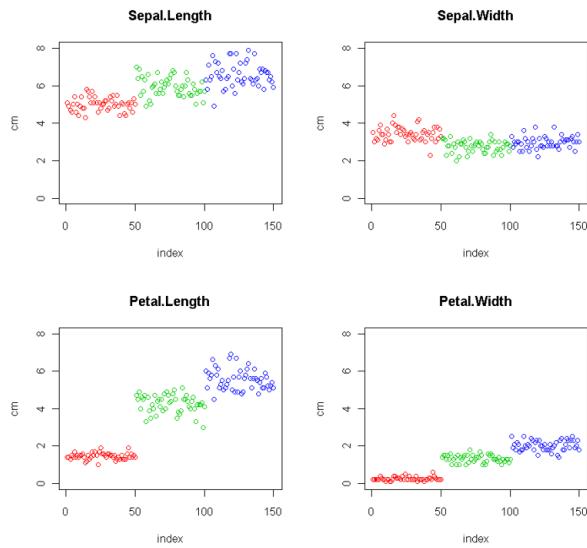
3.20 給定一函數 $f(x) = 12x^5 - 20x^3 + 7$,

- (a) 利用牛頓法 (Newton's Method) 求的所有解，迭代皆要 5 次以上。
- (b) 於閉區間 $[-1, 0.5]$ ，求出黎曼上和、黎曼下和，以及用 `integrate` 去計算此函數在此區間的面積。(提示：以 $n = 5$ 為例)

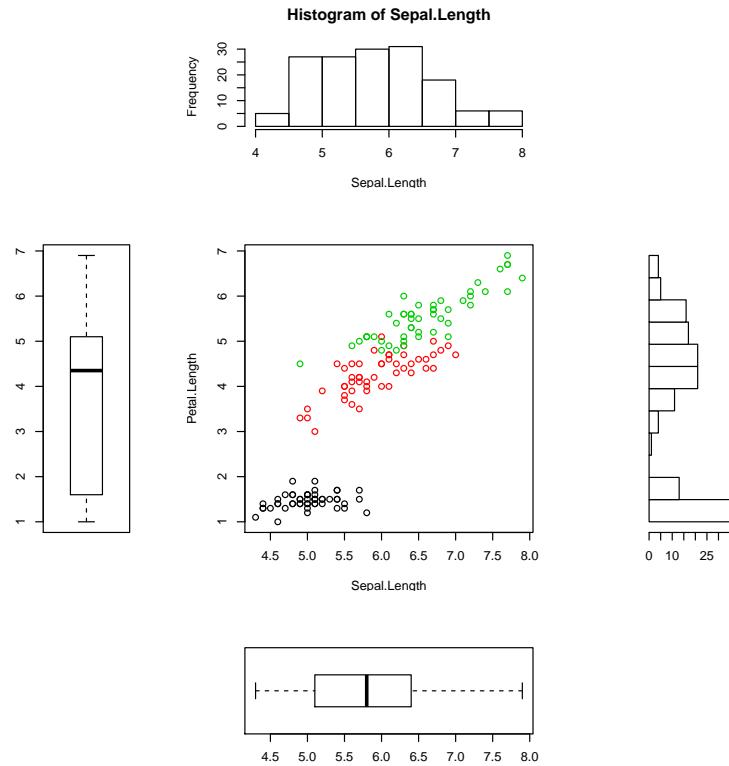
(c) 繪出下圖。



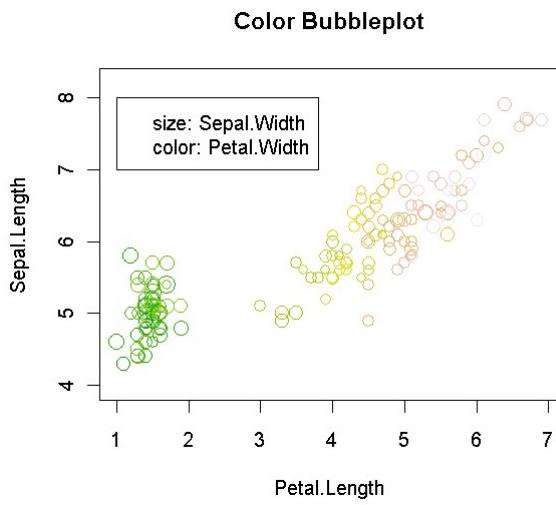
3.21 利用 `iris` 資料畫出下圖 (4 圖一頁，限用 `lapply`):



3.22 利用 `iris` 資料畫出下圖:



3.23 利用 `iris` 資料畫出下圖：

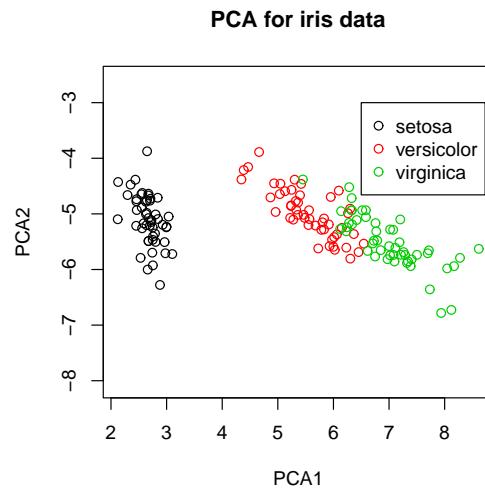


3.24 主成份分析法 (Principal Component Analysis, PCA) for Iris Data

- 用 R 函式 `cor` 求出 `iris` 四個連續型變數之相關係數矩陣 (命名為 `Sx2` 並印出)。
- 求出矩陣 `Sx2` 的特徵向量 (eigenvectors) 矩陣 (命名為 `evec` 並印出)。
- 資料第 K 個主成份的定義為原始資料矩陣與第 K 個特徵向量之乘積。請求出資料 `iris` 之第一及第二個主成分，分別命名為 `PCA1` 及 `PCA2`。(不用把

PCA1, PCA2 印出來)。

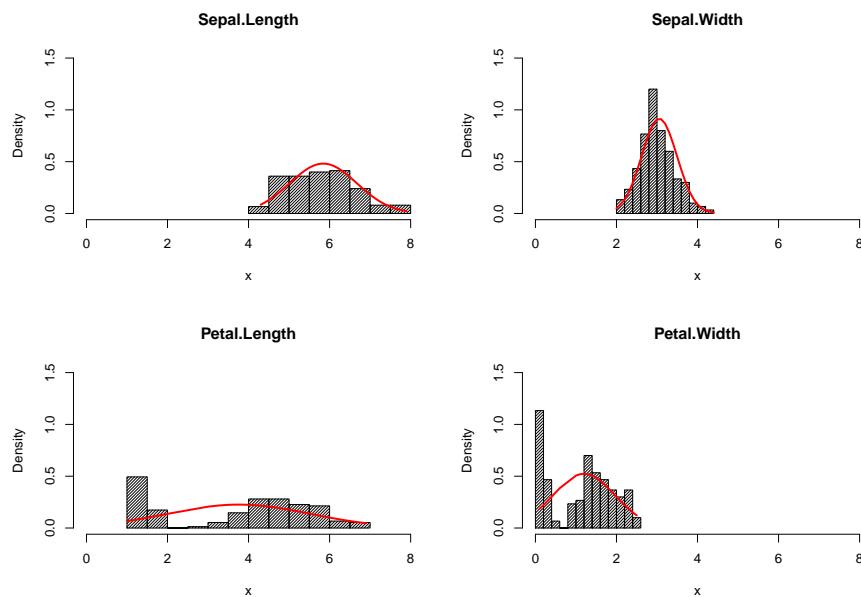
(d) 畫出 x -軸為 PCA1, y -軸為 PCA2 之散佈圖，並標上 Species 的顏色如下圖。



3.25 畫出 iris 資料 4 個連續變數之直方圖 (4 個圖一頁)，並各配適一常態機率機率曲線如下圖。常態機率密度函數為

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(註: (1) 不可用 `dnorm`; (2) 常態機率密度函數之參數以各變數之樣本平均數及樣本變異數為準， x 值即為資料點; (3) `mapply`)



3.26 (a) R 的內建資料 airquality 中，變數 "Ozone" 和 "Solar.R" 各有幾個 missing values?

- (b) 請選取完整且沒有 missing values 的資料，並存成 "airquality.complete"。
- (c) 在這完整的資料中，畫出 Month=5 時的 Wind (反應變數) 和 Temp (解釋變數) 的散佈圖，並加上一條迴歸線 (紅色)。
- (d) 在這完整的資料中，畫出 Month=5 和 7 時的 Wind (反應變數) 和 Temp (解釋變數) 的散佈圖，並加上一條迴歸線 (紅色)。其中 Month=5 和 7 要有不同的符號 (例如: a, b)，不同的顏色 (例如: blue, green)。要加 legend，要加 xlab, ylab, main。
- (e) 將最後一張圖存成 jpg 檔，然後再 MS Word 裡插入此圖形。

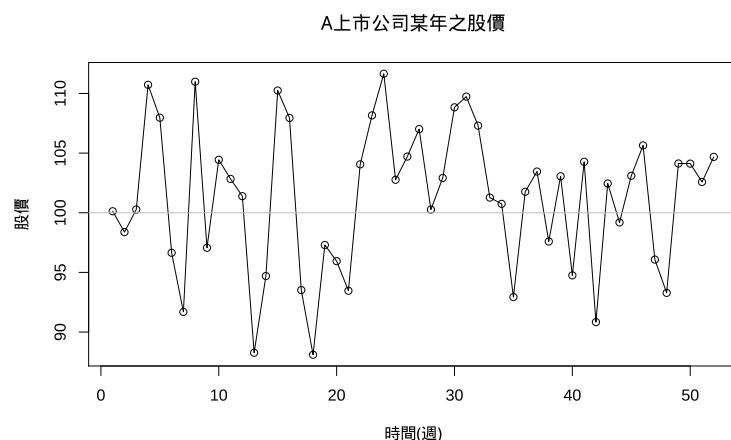
3.27 R 內建資料集 mtcars 是一汽車趨勢道路測試資料 (Motor Trend Car Road Tests) (1974 年)，資料包括 32 款汽車在油耗及 10 個汽車設計和性能測試相關的數據。11 個變數依序為: mpg (Miles/(US) gallon, 公哩/加侖), cyl (Number of cylinders, 氣缸數), disp (Displacement, 容量), hp (Gross horsepower, 總馬力), drat (Rear axle ratio, 後輪軸比) wt (Weight, 重量), qsec (1/4 mile time, ¼ 哩的時間), vs (Engine, 發動機類型), am (Transmission, 變速器) gear (Number of forward gears, 前進檔位數), carb (Number of carburetors, 化油器數)。利用此資料集畫出變數 mpg, disp, hp, drat, wt, qsec 之索引圖。(注意: 一頁 6 張圖 (2×3)，各圖上之符號點的顏色以 cyl 上色，各圖需加註「變數」標題)(提示: ? mtcars)

3.28 利用 R 內建之資料集 mtcars {datasets}，畫出變數 mpg, disp, hp, drat, wt, qsec 之 side-by-side 盒形圖。(每個變數之盒形需不同顏色，需加註標題)

3.29 A 上市公司某年 (52 週) 之股價 (stock.price) 如下所列。

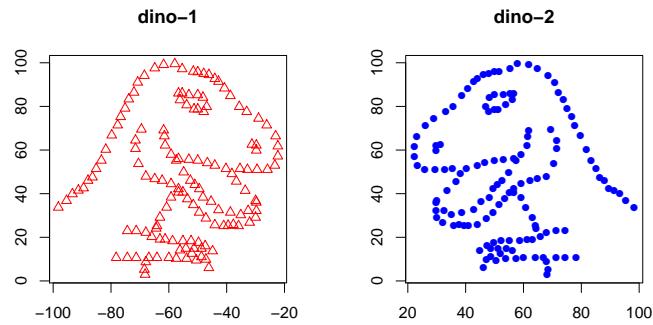
```
set.seed(12345)
time.points <- sample(1:52)
stock.price <- rnorm(52, 100, 5)
```

- (a) 畫出折線圖如下：

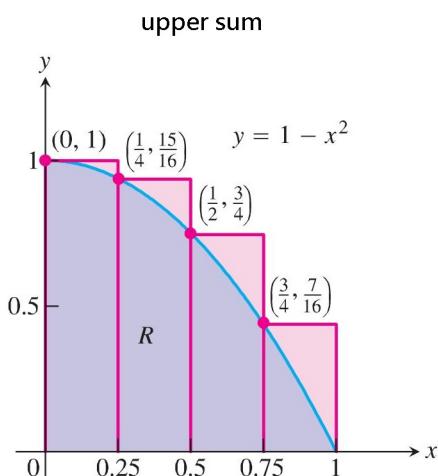


- (b) 將上小題之圖形以 R 指令 jpeg 及 pdf 存出 (各命名為 price.jpg 及 price.pdf)。
 (此題僅列出執行之程式碼，注意圖形檔之中文字需正常顯示)

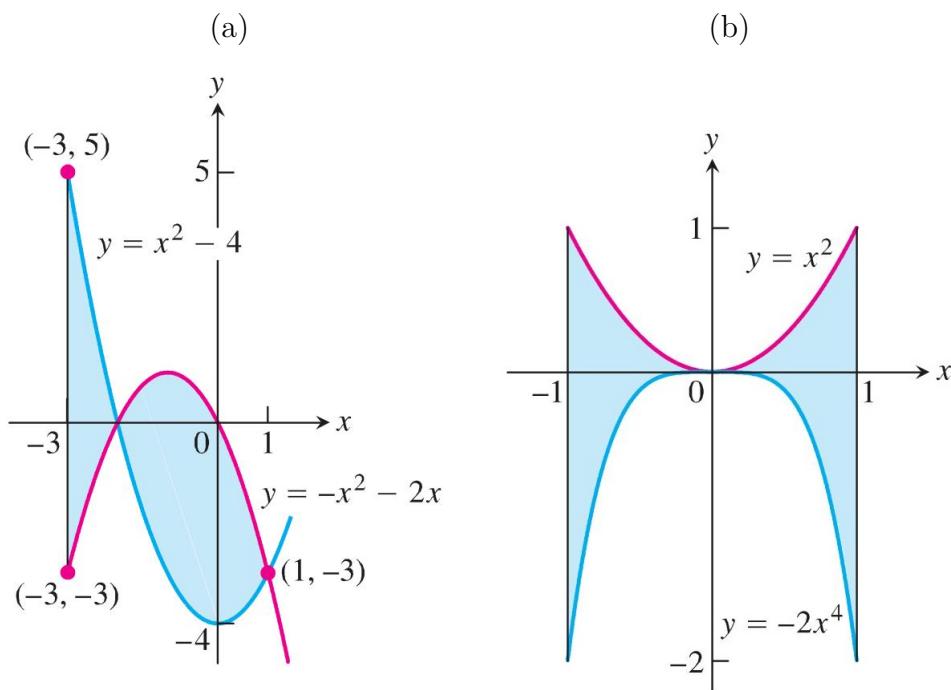
3.30 安裝”datasauRus” 套件，畫出下列圖形。(一頁兩張圖)



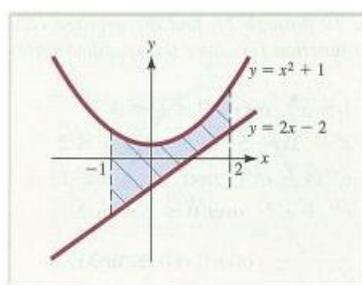
3.31 畫出下圖。



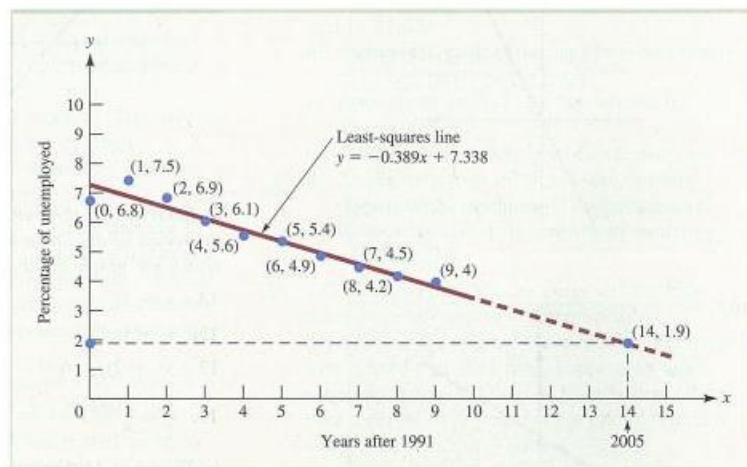
3.32 畫出下圖，需有淡藍色陰影。(提示: polygon, arrows, segments, lines, text)



3.33 畫出下圖。(提示: seq, function, plot, points, text, polygon, arrows, segments)



3.34 畫出下圖。

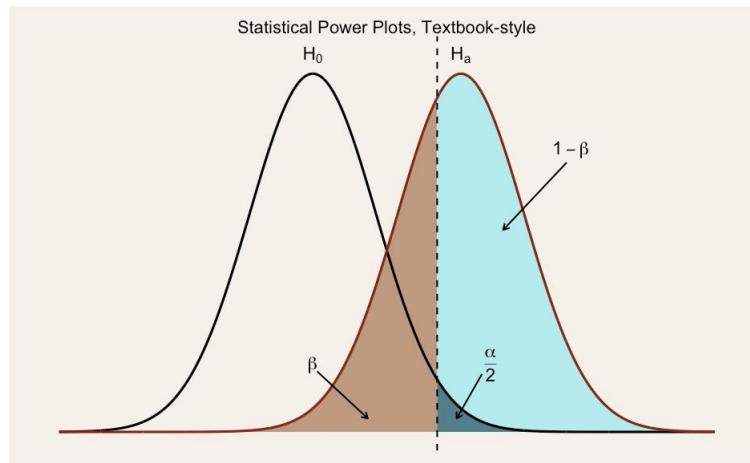


3.35 畫出下圖。在 H_0 條件下，一曲線為平均數 $\mu = 0$ ，標準差 $\sigma = 1.5$ 的常態機率密
R 練習題 (v2021.09) October 22, 2021

度函數; 在 H_a 條件下，另一曲線為平均數 $\mu = 3.5$ ，標準差 $\sigma = 1.5$ 的常態機率密度函數；常態機率密度函數為

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(註: (1) 不可用 `dnorm`, (2) 有虛線, (3) 曲線以下有顏色 (`polygon`))



3.36 (極座標) 以下是微積分課本 (Thomas' Calculus, Metric Edition; 12 edition, page 627) 有關極座標的定義:

Definition of Polar Coordinates

To demo polar coordinates, we first fix an origin O (called the pole) and an initial ray from O (Figure 11.18). Then each point P can be located by assigning to it a polar coordinate pair (r, θ) in which r gives the directed distance from O to P and θ gives the directed angle from the initial ray to ray OP .

(a) 直角座標轉換: 極座標 (r, θ) 轉換為直角座標 (x, y) 之公式如下:

$$x = r \cos \theta, \quad y = r \sin \theta$$

試寫一個 R 函式 (命名為 `polar2xy`)，輸入為 (r, θ) ，輸出為 (x, y) 。以 $(r, \theta) = c(2, \pi/3)$ 測試。

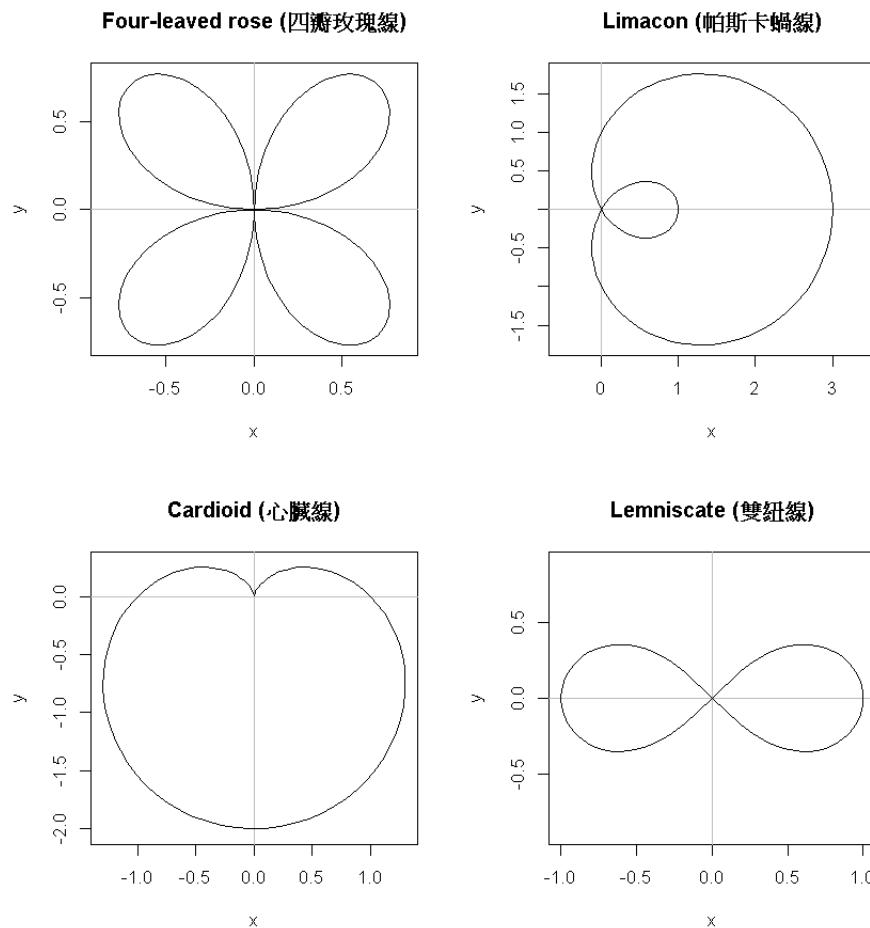
(b) 畫出下列四極座標圖形:

- i. Four-leaved rose (四瓣玫瑰線): $r = \sin 2\theta$.
- ii. Limaçon (帕斯卡蝸線): $r = 2 \cos \theta + 1$.
- iii. Cardioid (心臟線): $r = 1 - \sin \theta$.
- iv. Lemniscate (雙紐線): $r^2 = \cos 2\theta$.

(提示)

- 圖形之 x -axis, y -axis 之比例: `asp = 1`。
- 前三個圖之 θ 範圍為 $(0, 2\pi)$, R 程式碼為 `theta <- seq(0, 2*pi, length=100)`

- 雙紐線圖之 θ 範圍為 $(-\pi/4, \pi/4)$ 。解出 r 值之程式碼為
 $r <- c(sqrt(cos(2*theta)), -sqrt(cos(2*theta)))$ 。



3.37 畫出下圖。(提示: `draw.ellipse {plotrix}`)

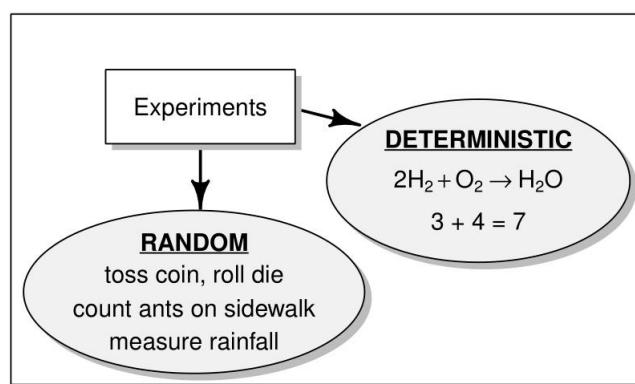
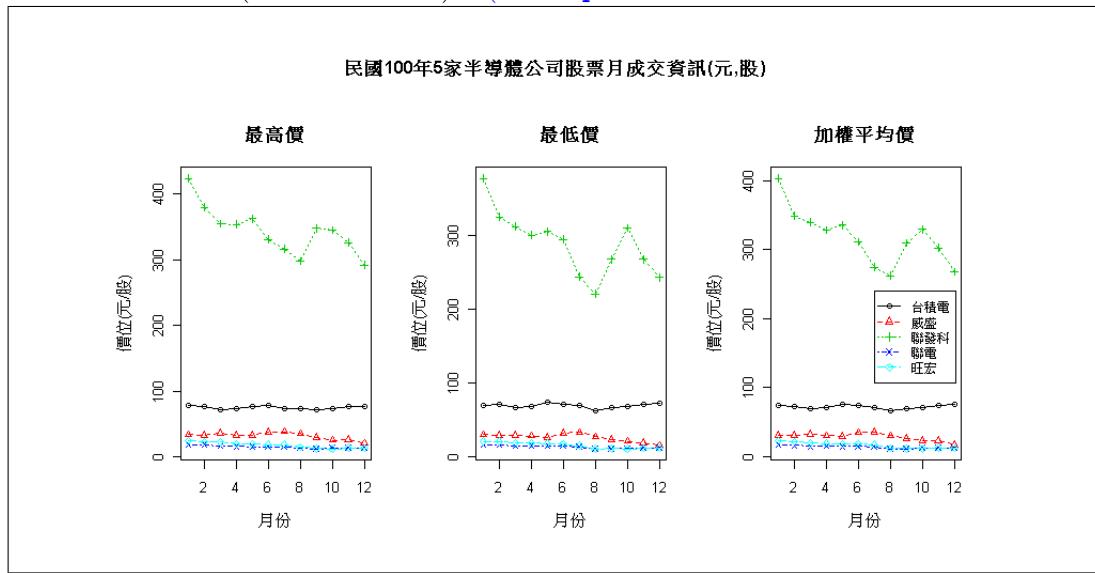


Figure 4.0.1: Two types of experiments

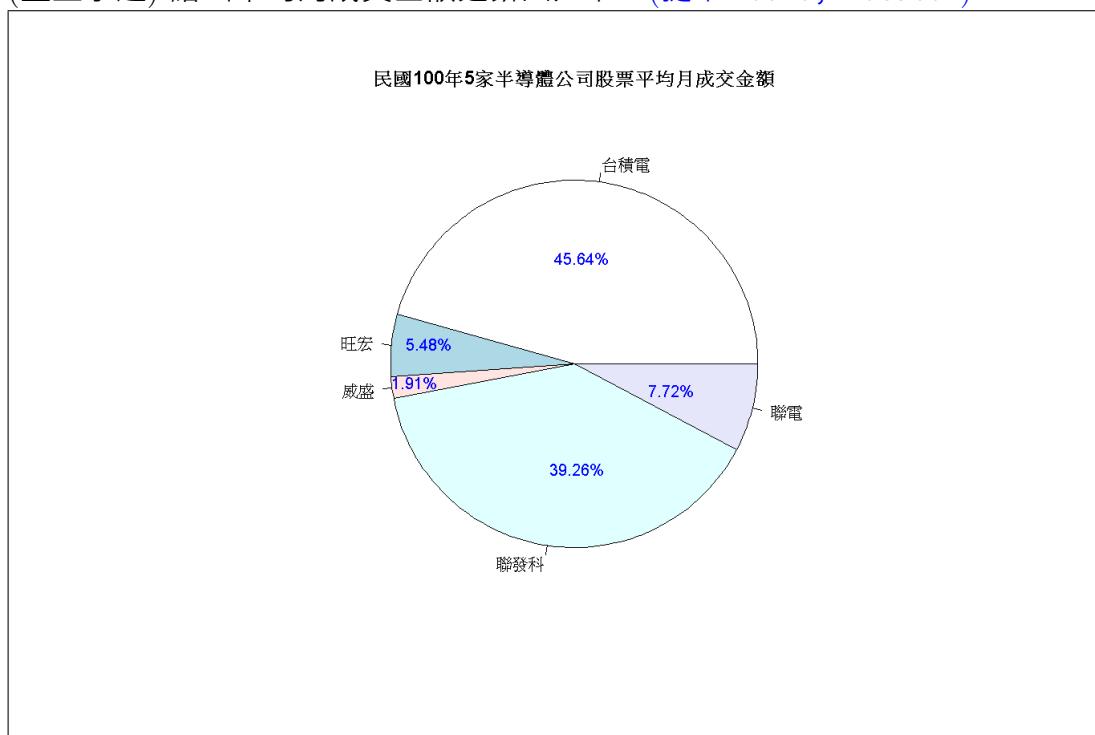
3.38 有民國 100 年 5 家半導體公司股票月成交資訊紀錄於資料檔 (`stock-data.txt`) 中，

- 請讀入此資料並列印出前 5 筆紀錄。

(b) 請繪出以下圖形 (一頁三個圖形)。(提示: `par(..., oma)`, `title(..., outer)`)



- (c) 請將資料中的「成交筆數」、「成交金額」及「成交股數」轉成數值型變數後，列印出此資料前 5 筆紀錄。
(提示: `as.numeric(gsub('\\,', ',', "100,578,274,926"))`)
- (d) 計算 5 家半導體公司股票之「平均」月成交金額。
- (e) (呈上小題) 繪出平均月成交金額之餅圖如下。(提示: `text`, `locator`)



3.39 資料集檔案 `WaterQuality_Hsinchu_modity.xlsx` 為「新竹市河川水質」近一年「監測資料」之紀錄，主要欄位為：河川名稱、測站名稱、採樣日期、採樣時間、溶氧量 (DO)、生化需氧量 (BOD5)、懸浮固體 (SS)、pH 值 (pH)、導電度 (EC)、水溫 (WT)、大腸桿菌群 (Coli_G)、氨氮 (NH3_N)、鉛 (Pb)、六價鉻 (Cr6)、鎘 (Cd)、銅 (Cu)、鋅 (Zn)、鎳 (Ni) 等等。請參見網址: <https://data.gov.tw/>

dataset/67604。

- (a) 讀取此資料集檔案，印出此資料集之紀錄筆數及欄位個數，並印出所有欄位名稱。
- (b) 產生一具有三個類別的因子（順序）變數：水溫組（命名為 WT.group）。規則如下：若水溫 (WT)(°C) 高於平均的 1 倍標準差，則 WT.group 為高溫組，若水溫 (WT)(°C) 低於平均的 1 倍標準差，則 WT.group 為低溫組，其餘為均溫組。印出三個類別的個數（提示：ifelse, table）
- (c) 畫出下列四個變數之索引圖：溶氧量、pH 值、大腸桿菌群、氨氮。每一張圖之點符號（symbol）顏色為水溫組之類別。（註：一頁 4 張圖，每張圖的標題為變數名稱及其單位）
- (d) 畫出下列四個變數之直方圖：溶氧量、pH 值、大腸桿菌群、氨氮。（註：一頁 4 張圖，每張圖的標題為變數名稱及其單位）

3.40 教學助理教學評鑑資料 (Teaching Assistant Evaluation Data Set⁴) (檔案: tae.data): 威斯康辛大學麥迪遜分校統計系針對 151 位教學助理，實施教學評鑑，為期三個學期及二個暑期，並將評鑑的結果分為 ("low", "medium", and "high") 三個等級。以下為此資料欄位的資訊 (Attribute Information):

NativeEng: Whether or not the TA is a native English speaker (binary)
 1=English speaker, 2=non-English speaker
 Instructor: Course instructor (categorical, 25 categories)
 Course: Course (categorical, 26 categories)
 Semester: Summer or regular semester (binary) 1=Summer, 2=Regular
 ClassSize: Class size (numerical)
 Scores: Class attribute (categorical) 1=Low, 2=Medium, 3=High

今有一同學想藉由一些統計圖來了解資料中的四個變數「NativeEng, Semester, ClassSize, Scores」的分佈及它們之間的相關，請你幫幫他。（提示：儘可能將所有有助於了解資料的基本統計圖畫出。一頁多張圖有助於做比較。要注意尺度，要註明圖的標題）

3.41 有某班學生之微積分成績明細紀錄於資料檔 (score2015.txt) 中，其中成績以 60 分為及格，100 分為滿分，成績空白以零分計。學期總成績計算方法如下：(i) 配分比例為：小考成績佔 40% (各次小考平均配分)、期中考佔 30%、期末考佔 30%。助教實習課及出席次數不算分。

- (a) 讀取資料檔，並印出前 5 筆成績紀錄。
- (b) 計算學期總成績，並畫出學期總成績之直方圖 (hist)。（請加上合適之主標題及 x-, y- 軸之標號）

⁴<https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation>

- (c) 畫出期中考與期末考之散佈圖， x -軸及 y -軸之範圍皆為 $(0 \cdot 100)$ ，加上一通過原點之 45 度直線 (灰色)，散佈圖上之符號點以兩種顏色 (紅、綠) 區分性別 (男、女)。(請加上合適之主標題)

3.42 (**heatmap 指令練習**) mydata 紀錄某班 15 位學生六次考試的成績，試別 (colID) 分為口試及筆試 (各三次)。學生依志願分為三組 ("A"，"B"，"C") 測驗 (rowID)。請利用 heatmap 畫出此資料的熱圖，其中 ColSideColors 的顏色是以 colID 為依據，而 RowSideColors 的顏色是以 rowID 為依據。

```
> mydata <- data.frame(matrix(sample(0:100, 15*6, replace=T), ncol=6))
> rownames(mydata) <- paste0("student", 1:15)
> mydata
    X1 X2 X3 X4 X5 X6
student1 44 94 63 60 95 89
student2 75 48 23 30 59 37
student3 5 32 93 96 70 74
student4 35 22 49 99 58 19
student5 4 8 46 15 25 23
student6 39 83 4 24 64 62
student7 20 55 28 46 49 98
student8 54 35 54 86 79 49
student9 14 45 7 49 68 42
student10 30 72 59 79 36 91
student11 42 47 62 45 9 48
student12 68 83 78 5 81 72
student13 13 94 13 58 16 93
student14 33 89 89 8 97 8
student15 92 83 28 48 6 42
> colID <- rep(c("oral", "written"), each =3)
> colID
[1] "oral"     "oral"     "oral"     "written"  "written"  "written"
> rowID <- sample(c("A", "B", "C"), 15, replace=T)
> rowID
[1] "B" "B" "B" "B" "B" "B" "C" "B" "B" "A" "C" "B" "C" "C" "C"
```

3.43 (**scatter3D {plot3D} 指令練習**)

- (a) 安裝 dimRed 套件，並載入到 RStudio。利用指令 dataSetList 列出此套件內建之資料集名稱。
- (b) 利用指令 loadDataSet 分別產生 "3D S Curve"、"FishBowl"、"Helix" 三個資料集各 1000 個資料點。以 scatter3D {plot3D} 畫出三種不同角度之散佈

圖 (資料點以彩虹顏色呈現)(角度自選) (每一資料集為一頁三張圖 (即三個不同角度, 一列三欄))。

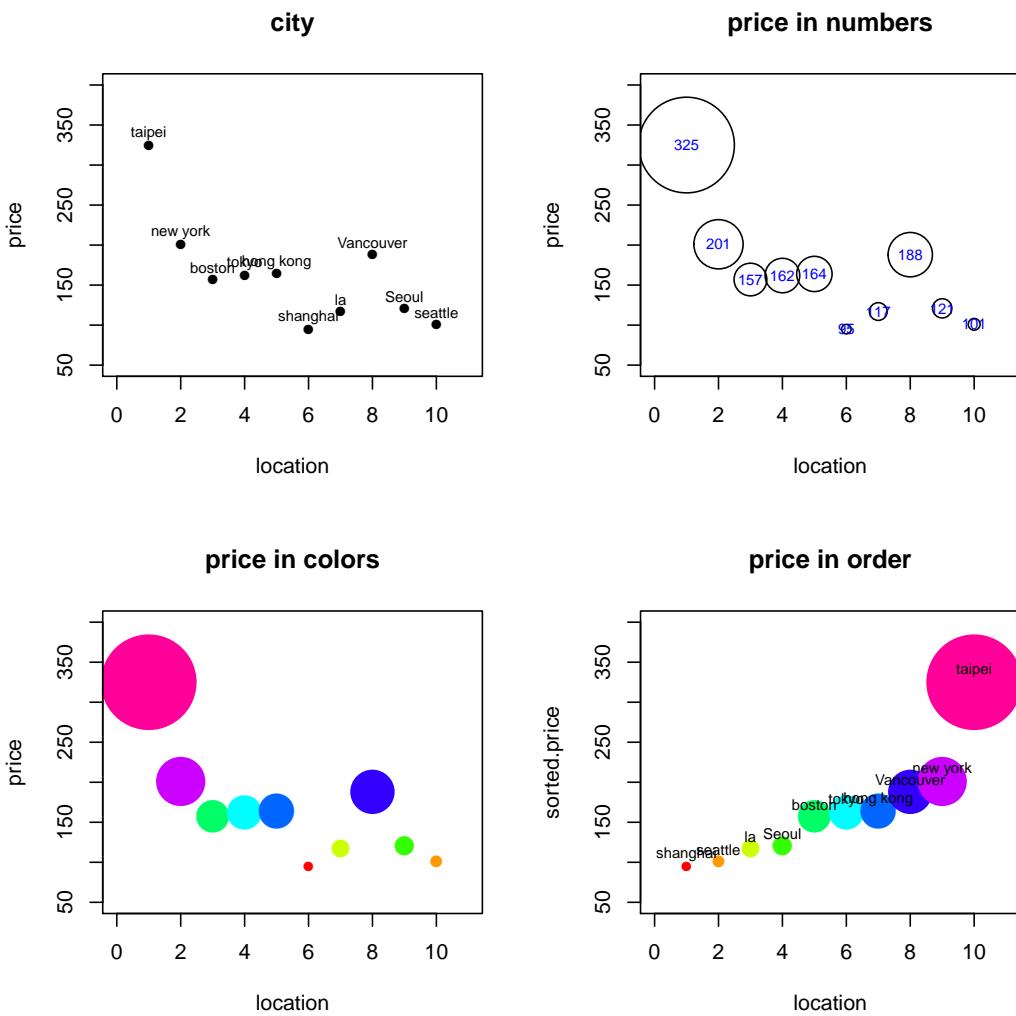
3.44 讀取資料檔 city.txt:

- (a) 依照下列公式，將 price 之值轉換為範圍介於 (1~10) 之值，並列印出來。

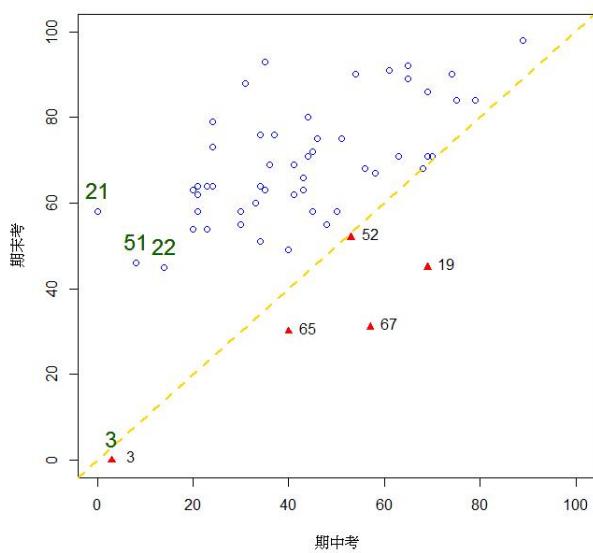
$$\text{transformed.price} = 9 \times \frac{\text{price} - \min(\text{price})}{\max(\text{price}) - \min(\text{price})} + 1$$

- (b) 繪出下列一頁 4 圖 (其中 Bubble plot 之泡泡大小是依據上小題)。

(提示: sort, order, rainbow(10))



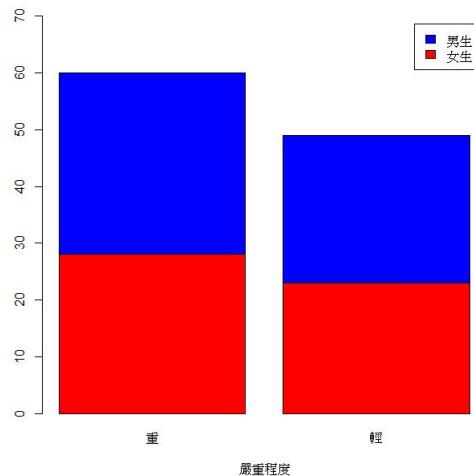
3.45 「score01.csv」為記錄某班的統計學期中及期末成績資料檔，繪出期中考與期末考的散佈圖，其中橫軸為期中考，縱軸為期末考。圖中以「紅色三角點」符號表示其期末成績退步者，並在此符號旁邊標示學號。另外，以「綠色且字型較大之數字」表示學號，標示在學期成績小於 35 分之同學。



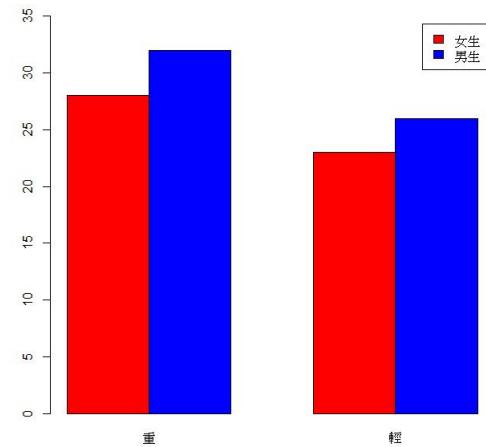
3.46 「diabetes.csv」為一記錄糖尿病的資料集，其變數解釋如下：

- 編號 (id)
 - 性別 (sex)： 1=女生 2=男生
 - 年齡 (age)
 - 教育程度 (edu)： 1=不識字， 2=小學， 3=國(初)中， 4=高中(職)， 5=大專以上
 - 糖化血紅素 (a1c)
 - 體重 (wt)
 - 膽固醇 (ldl)
 - 收縮壓 (sbp)
 - 舒張壓 (dbp)
 - 嚴重度 (a1cgp)：1=輕 2=重
- (a) 讀入資料集「diabetes.csv」(命名為 diabetes)。判斷出哪些變數中含有 NA?
 - (b) 判斷變數 sex, edu, a1cgp 是否為 factor 類別? 若不是，請轉換成 factor 類別變數。
 - (c) 刪除具有 NA 的紀錄，並將剩餘之完整資料另存成一 R 資料框 (命名為 diabetes.c)。印出前 10 筆紀錄。
 - (d) 繪出「教育程度」之長條圖。(每一長條需不同顏色)
 - (e) 繪出下列統計圖。

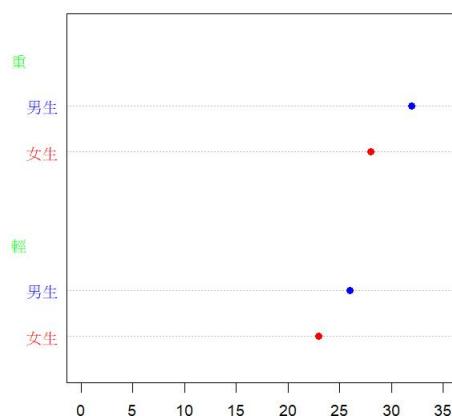
(a)



(b)

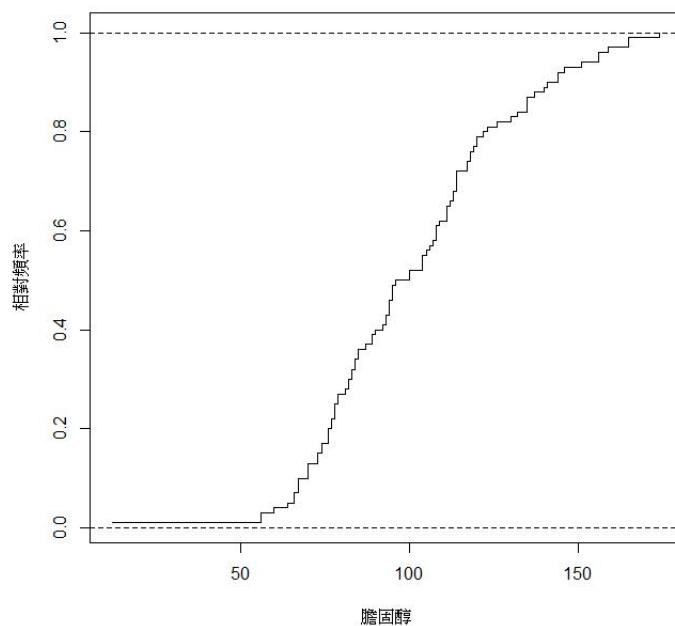


(c)



(f) 畫出「膽固醇」之直方圖。(需加標題)

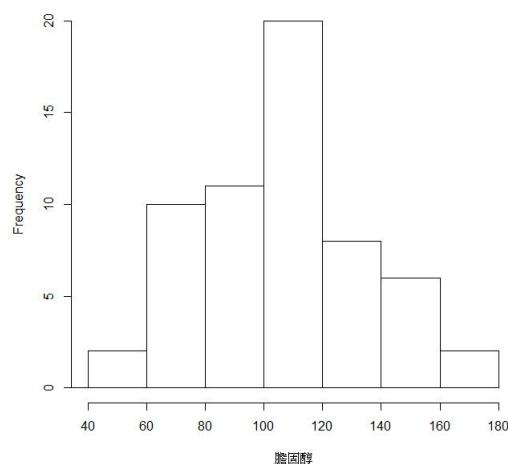
(g) 列出「膽固醇」之次數分配表，依此畫出累積機率分配圖。



(h) 繪出下列統計圖。

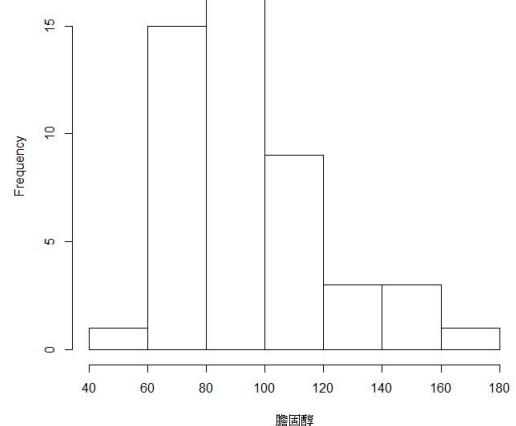
(a)

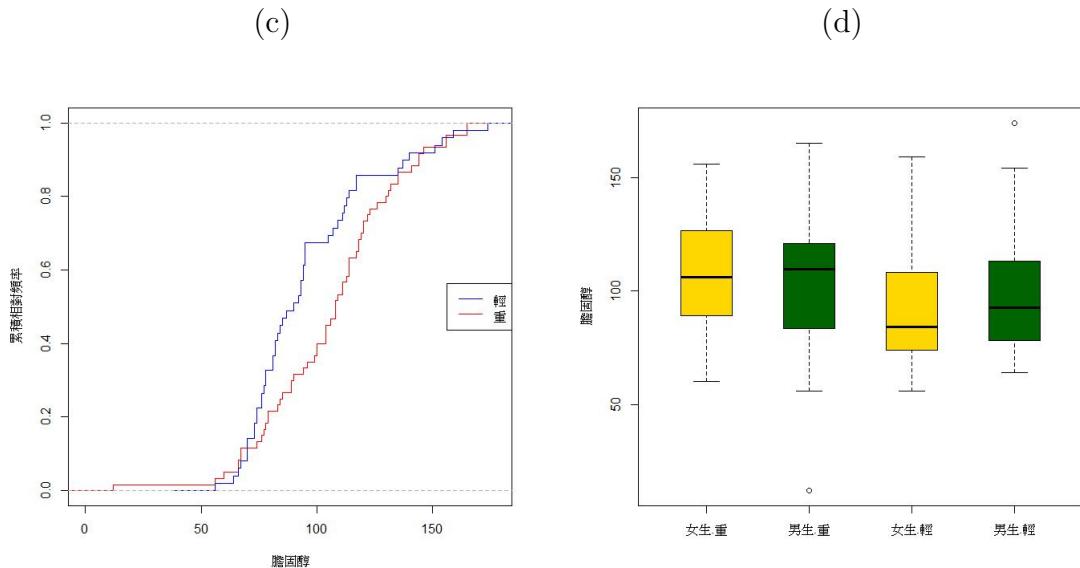
嚴重程度=重



(b)

嚴重程度=輕

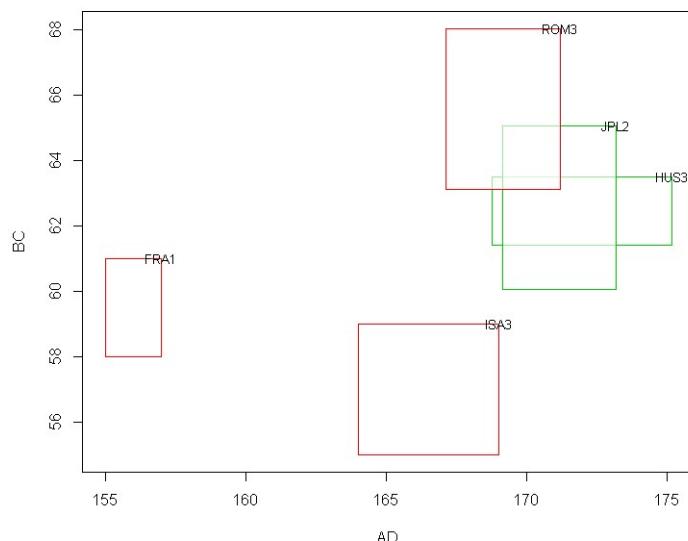




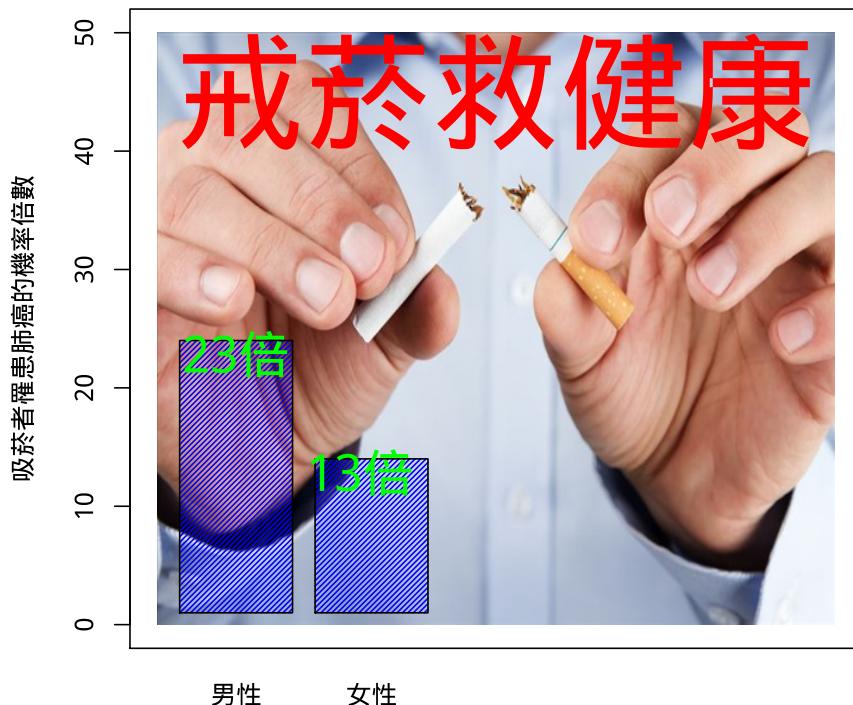
3.47 於座標平面上，以不同顏色繪出下列各組之「三角形」(亦即將三個座標點連線)，且「三角形」旁需標記 A E。(註：若三個座標點無法形成三角形也需連線)

1. A: (4, 6) (-2, 8) (-8, 10)
2. B: (16, 14) (8, 2) (2, 6)
3. C: (0, 4) (8, 4) (4, $4+4\sqrt{3}$)
4. D: (-4, 2) (4, 6) (-2, 8)
5. E: (12, 9) (8, 2) (2, 1)

3.48 有一區間 (min, max) 資料紀錄檔: intervals.txt，共 5 個觀察值 (分為兩群) 及兩個區間變數 AD 和 BC，畫出散佈圖如下。



3.49 (統計圖 + 背景) 畫出下圖 (背景圖片檔名”20141230134054104_o.jpg”)。(圖片來源:<https://www.cmoney.tw/notes/note-detail.aspx?nid=22721>)



3.50 母親節快到了，真妮佛 同學想要利用 R 程式畫出下圖表達對媽媽的養育之恩，請大家幫她畫一下吧！

(提示)

(a) 心形之方程式如下：

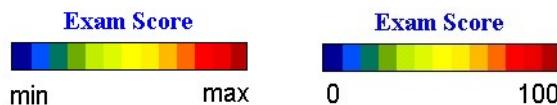
$$\begin{aligned}
 x(\theta) &= 16 \sin^3(\theta), \\
 y(\theta) &= 13 \cos(\theta) - 5 \cos(2\theta) - 2 \cos(3\theta) - \cos(4\theta), \\
 -\pi &\leq \theta \leq \pi.
 \end{aligned}$$

(b) 會利用到的指令及參數: `plot(..., xaxt="n", frame.plot=F)`, `polygon`, `text`



3.51 資料: student-score.txt。

畫出此資料的 heatmap(列及行皆不排序) · 依照以下兩個彩虹色階 (其中 min, max 為資料中的最小值及最大值) · 各畫出 (a) Range Matrix Condition, (b) Range Column Condition, (c) Range Row Condition 的 heatmap。(共 6 個圖畫在同一頁 · 每個子圖需有標題)



3.52 (直方圖) 繪製直方圖有兩個最重要的參數: 帶起始值 (Bin origin) 及帶寬 (Bin widths)。利用 R 內建之資料集 (CO2 {datasets}) · 依下列各條件畫出變數 uptake 之直方圖 · 並使用 RColorBrewer 套件之 Spectral 色階 · 為每一帶寬上色。(注意: 一頁 4 張圖 (1×4)) · 且每個圖形需加上適合之標題。)

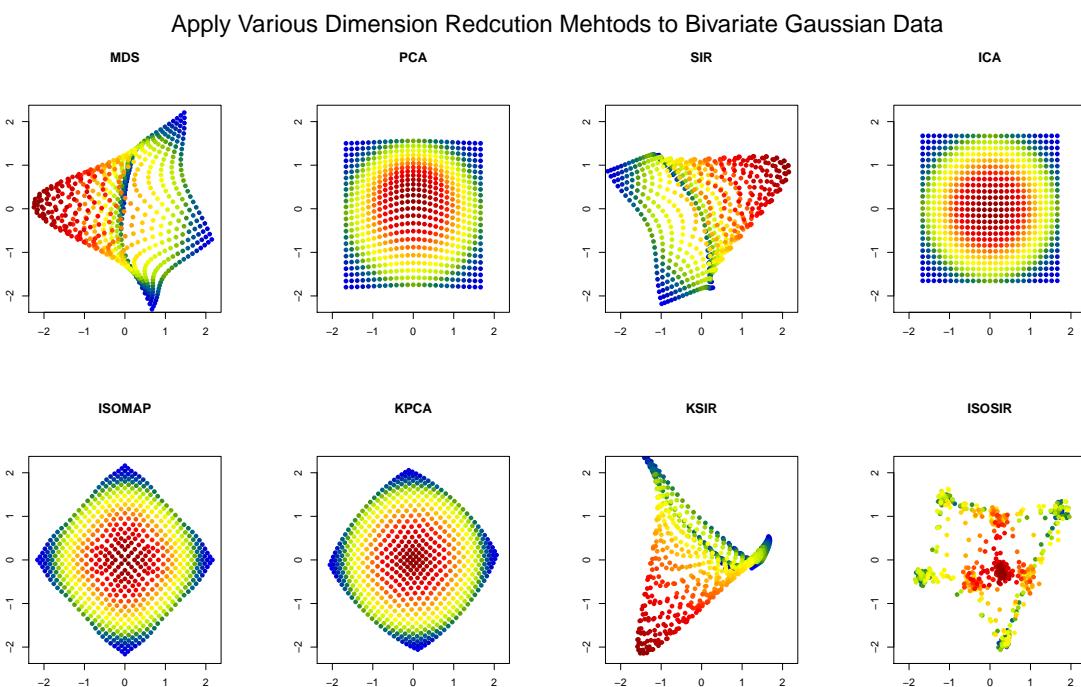
- (a) 帶起始值為 0 · 帶寬為 5。
- (b) 帶起始值為 5 · 帶寬為 5。
- (c) 帶起始值為 5 · 帶寬為 10。
- (d) 帶起始值 0 · 帶終點值為 50 · 共 10 組帶寬。

3.53 利用 R 內建之資料集 (CO2 {datasets}) · 畫出 conc 及 uptake 索引圖。(二圖一頁)(並以變數 Plant 之類別上色)

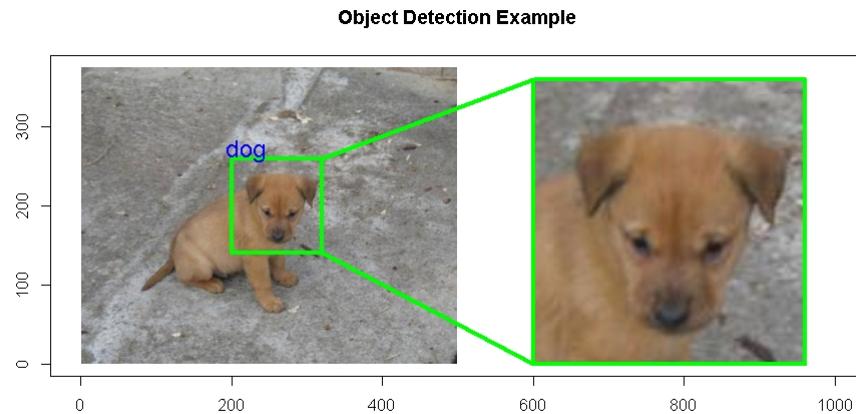
3.54 利用二元常態分佈資料檔 GaussianDataDR.csv 畫出下圖 · 其中各個子圖之橫座標為資料中的 (x, y, z) 子資料集經由維度縮減後的第一個方向 (例如: MDS.1) · 而子圖中之縱座標為維度縮減後的第二個方向 (例如: MDS.2)。資料中各欄位所代表之意義如下:

- color: 資料點之顏色。
- x, y , z: $z = f(x, y)$ · 其中 f 為二元標準常態機率密度函數。
- MDS.1, MDS.2: MDS 為多元尺度法 (multidimensional scaling) 之簡稱，是一種維度縮減方法。MDS.1 代表資料經由 MDS 維度縮減後所得到的第一個方向，MDS.2 代表資料經由 MDS 維度縮減後所得到的第二個方向。其餘欄位為不同的維度縮減方法：
 - PCA: 主成分分析 (Principal Component Analysis)
 - SIR: 切片逆迴歸法 (Sliced Inverse Regression)
 - ICA: 獨立成份分析 (Independent Components Analysis)
 - ISOMAP: 等軸距特徵映射 (Isometric Feature Mapping)
 - KPCA: 核化主成分分析 (Kernal Principal Component Analysis)
 - KSIR: 核化切片逆迴歸法 (Kernal Sliced Inverse Regression)
 - ISOSIR: 等軸距切片逆迴歸法 (Isometric Sliced Inverse Regression)

維度縮減方法可參照 wikipedia 說明: https://en.wikipedia.org/wiki/Dimensionality_reduction



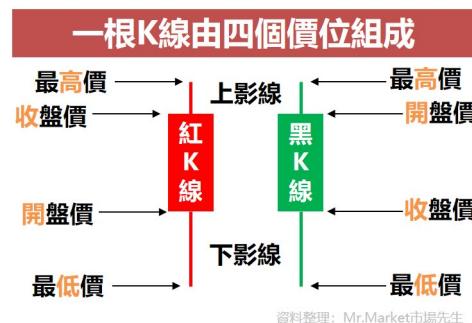
3.55 模擬影像 Object Detection。利用圖檔 dog1257.jpg，畫出下圖。(提示: plot, rasterImage, rect, segments, text, dog_subset <- dog[200:320, dims[2]-260:140,])



3.56 (K 線圖) 資料檔 IBM201701.csv 為 IBM 公司於 2017 年一月之股價資訊，包含開盤價 (Open)、最高價 (High)、最低價 (Low)、收盤價 (Close)、交易量 (Volume) 及修正指數 (Adjusted)。一般常見的技術分析是 K 線圖，它是根據股價一天 (或者某一周期) 走勢中形成的四個價位：開盤價、最高價、最低價、收盤價 (開、高、低、收) 繪製而成。細節及圖片請參照<https://rich01.com/what-is-k-bar-charts>

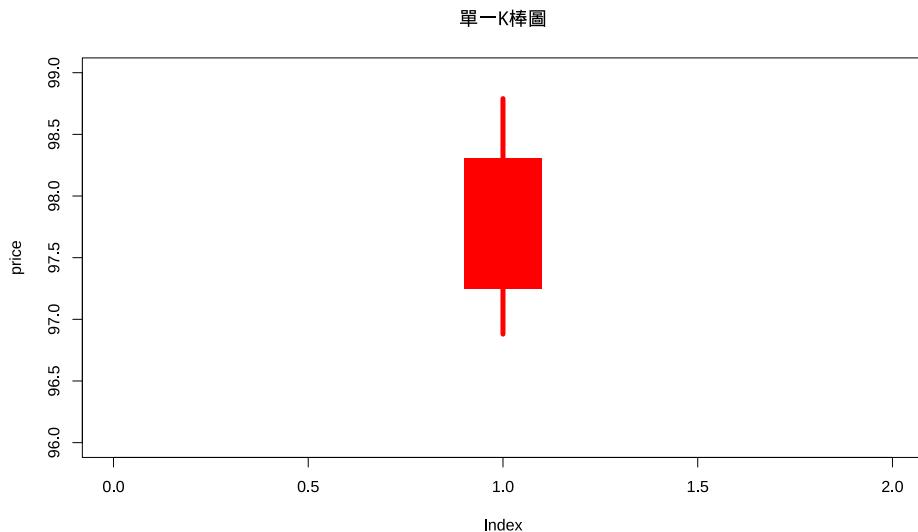
「K線因為長的像蠟燭，也有人稱為蠟燭線或蠟燭圖，也有人稱為K棒，或K棒圖。這一根蠟燭本身代表開盤價、收盤價，而蠟燭兩端燭芯代表最高價、最低價。投資人可以透過蠟燭本身的長度，來判斷股票當天的漲跌程度，但不論是紅K線或黑K線，最高價永遠在最上方、最低價永遠在最下方。」

- * 收盤價>開盤價：代表股價上漲，會以紅色來表示，稱為紅K線、陽線。
- * 收盤價<開盤價：代表股價下跌，會以綠色來表示，稱為黑K線、陰線。
- * 收盤價=開盤價相同：稱為十字線。
- * 最高價、最低價，分別是細細的上影線及下影線。

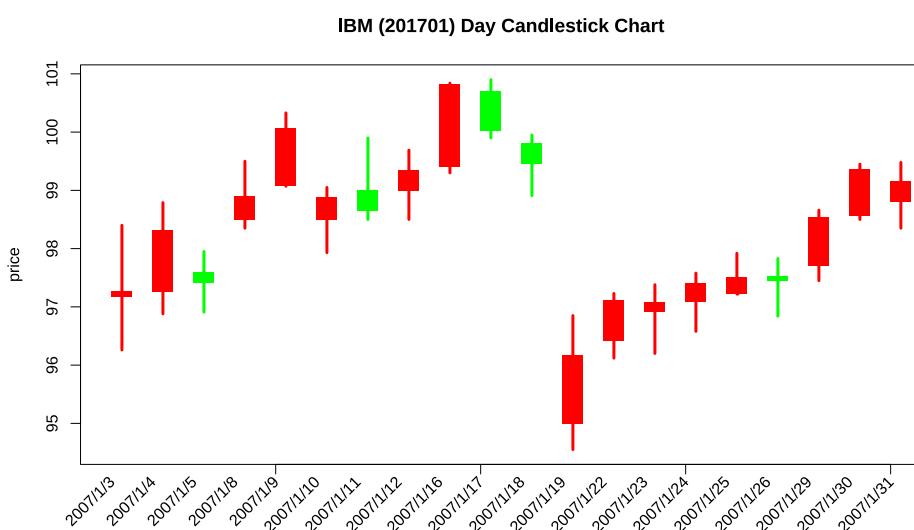


- (a) 讀入資料檔 IBM201701.csv，並列印其結構及全部資料。
- (b) IBM201701.csv 中的第二筆紀錄 (2007/1/4) 為: (Open, High, Low, Close) = (97.25, 98.790001, 96.879997, 98.309998)。今欲畫出此單一 K 棒圖，步驟如下:
 - 設定 4 個變數值: Open, High, Low, Close。

- 產生一個空白底圖，設定好 `xlim`、`ylim`、`y` 軸標號和主標題。
- 以 2 次 R 指令 `ifelse` 來判別 K 棒顏色：當「收盤價 > 開盤價，K 棒為紅色」；當「收盤價 < 開盤價，K 棒為綠色」；當「收盤價 = 開盤價，K 棒為灰色」。
- 主體部份由 R 指令 `rect` 畫出。(找出起始座標和終點座標，`border = NA`，寬度自訂，並查指令？`rect`)
- 上下影線則由 2 次 R 指令 `segments` 畫出。(找出起始座標和終點座標，`lwd = 5`，並查指令？`segments`)



(c) 試寫一個 R 函式 (命名為 `Candlestick_Chart`)，輸入為資料檔名 (格式為固定)、輸出 K 線圖。以資料檔 `IBM201701.csv` 測試。(提示：`ifelse`、`rect`、`axis`、`text`、`segments`)



3.57 (熱圖) 資料檔 `score1032.txt` 為某班之學期成績，其中「英文」及「語表」為文科科目，其餘為理科科目。繪製此成績資料之熱圖 (heatmap) 兩張：(a) 行列皆未

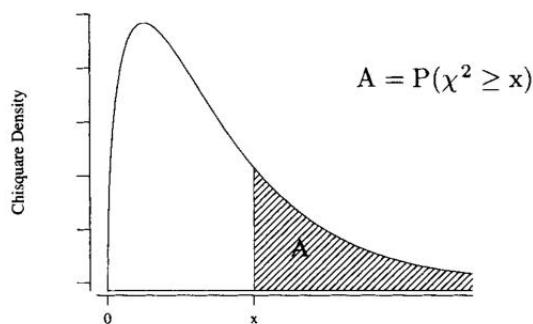
排序; (b) 行列皆適當排序 (自行選排序方法); 其它要求如下:

- 資料矩陣使用之色階為 `tim.colors {fields}`。
- 熱圖之「欄位」需有科目名，「文科」及「理科」以兩種顏色標在欄位之上。
- 熱圖之「列位」需有學號。「性別」則以兩種顏色標註在熱圖旁。

3.58 下列圖表截取自某一統計教科書之附錄。(a) 畫出卡方分佈之機率密度函數圖 (如下圖，要求: 需自選合適之參數，需有座標軸之標號、曲線下面積陰影、圖上之文字) 及 (b) 印出累積機率值表格 (如下表)。

570

Appendix D

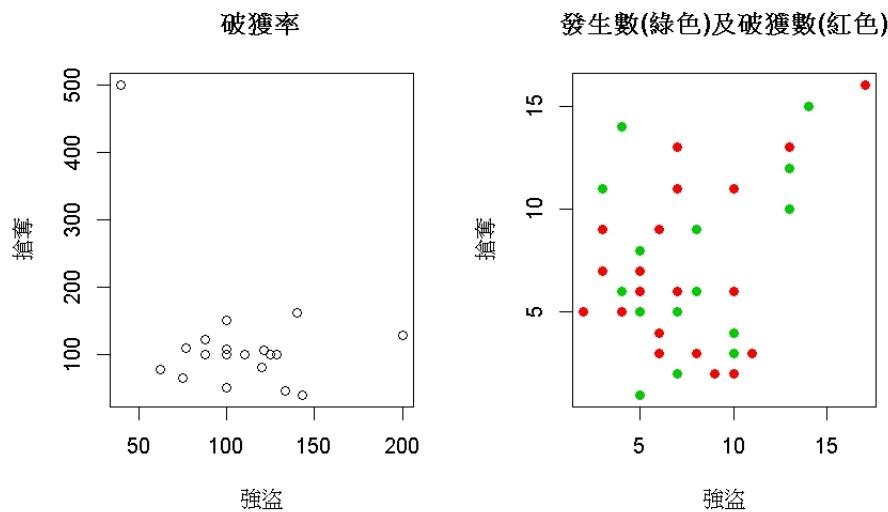


CHI-SQUARE DISTRIBUTION: The column headings are values for A , and the body of the table gives values for X .

df	0.995	0.990	0.975	0.950	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	15.987	18.307	20.483	23.209	25.188

3.59 讀取資料檔「104 年-即時犯罪資料統計數據.xls」，

- 列印出資料前 4 筆、後 4 筆紀錄及各變數所屬類別。
- 印出變數「案類別」各類別 (破獲率、發生數、破獲數) 之紀錄次數。
- 印出「案類別」為「破獲率」之列號如下:
1 7 11 12 15 18 21 23 26 28 31 34 37 41 45 46 47 49 50 56
- 畫出兩變數「強盜 vs. 搶奪」之散佈圖如下 (其中左圖案類別之發生數為「紅色」、「破獲數」為綠色))。

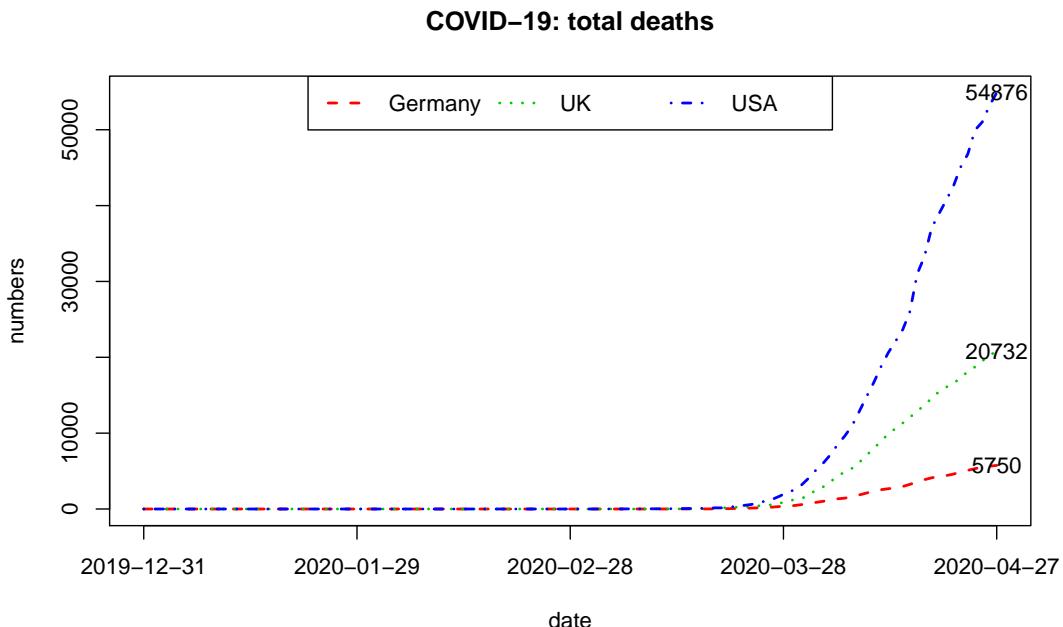


3.60 資料來源: <https://github.com/owid/covid-19-data/tree/master/public/data>

- 讀取「新冠肺炎」資料 (檔案: `owid-covid-data.csv`)，並印出前 3 筆及後 3 筆紀錄。
- 選取三個國家 (Germany, United Kingdom 及 United States) 其中兩變數 (日期 (`date`) 及總死亡人數 (`total_deaths`)) 之資料，存成一 R 資料框 (`data frame`) 類別物件，依日期，印出前 6 筆及後 6 筆紀錄如下。(提示: `==`, `data.frame`)

```
> head(mydata)
  date Germany UK USA
1 2019-12-31      0  0   0
2 2020-01-01      0  0   0
3 2020-01-02      0  0   0
4 2020-01-03      0  0   0
5 2020-01-04      0  0   0
6 2020-01-05      0  0   0
> tail(mydata)
  date Germany     UK     USA
114 2020-04-22 4879 17337 45063
115 2020-04-23 5094 18100 46784
116 2020-04-24 5321 18738 49963
117 2020-04-25 5500 19506 51017
118 2020-04-26 5640 20319 53189
119 2020-04-27 5750 20732 54876
```

(c) 依上小題之資料，畫出下圖。(提示：橫軸之日期標記僅選畫 5 天)



3.61 (perp 指令練習) 雙變量 (X_1, X_2) 常態分佈機率密度函數定義如下: Two random variables X_1 and X_2 are said to have a bivariate normal distribution with parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$, and ρ , if their joint probability density function is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right]$$

where

$$z = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}$$

and

$$\rho = \text{corr}(x_1, x_2) = \frac{\text{cov}_{12}}{\sigma_1\sigma_2}.$$

is the correlation of X_1 and X_2 and cov_{12} is the covariance of X_1 and X_2 . 試寫一雙變量常態分佈機率密度函數之 R 函式。輸入為 $(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \text{cov}_{12})$ ，輸出為 $f(x_1, x_2)$.

3.62 (承上題) 依照下列參數，畫出雙變量常態分佈機率密度函數圖。

(可參考: http://tagteam.harvard.edu/hub_feeds/1981/feed_items/177468)

- (a) $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \text{cov}_{12} = 0$.
- (b) $\mu_1 = \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 9, \text{cov}_{12} = 0$.
- (c) $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \text{cov}_{12} = 0.99$.

3.63 二元常態分佈記做 $(X_1, X_2) \sim BVN(\vec{\mu}, \Sigma)$ · 其聯合機率密度函數表示如下:

$$f(\vec{x}) = f(x_1, x_2) = \frac{1}{2\pi} \left[\frac{1}{\det(\Sigma)} \right]^{1/2} \exp \left[\frac{-1}{2} (\vec{x} - \vec{\mu})^t \Sigma^{-1} (\vec{x} - \vec{\mu}) \right],$$

其中 $\vec{\mu} = (\mu_1, \mu_2)^t$ 為 (X_1, X_2) 之平均向量 · $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$ 為共變異數矩陣,
 $\det(\Sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$ 為共變異數矩陣之行列式。若假設有二元常態隨機變數
 (X_1, X_2) · 其平均向量為 $\vec{\mu} = (0, 1)^t$, 共變異數矩陣為 $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$ · 試回答下
 列各小題。

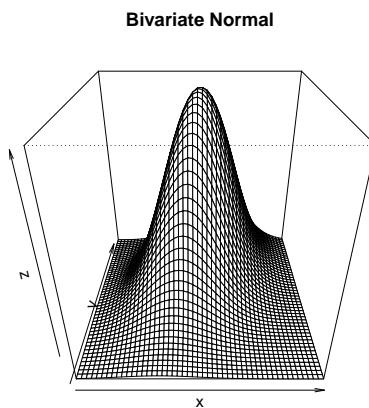
- (a) 利用 `dmvnorm` {mvtnorm} · 計算二元常態分佈 $f(x_1 = 1, x_2 = 2)$ 之機率密度函數值。
- (b) 利用 `dmvnorm` {mvtnorm} 及下列提示 · 計算二元常態分佈 $f(x_1 = -1, x_2 = 2), f(x_1 = 0, x_2 = 1), f(x_1 = 1, x_2 = 2)$ 之機率密度函數值。(提示: (1) `x1 <- c(-1, 0, 1)`; (2) `x2 <- c(2, 1, 2)`; (3) `cbind(x1, x2)`)
- (c) 利用 `persp` {graphics}, 畫出此二元常態聯合機率密度函數圖如下 (提示:
`outer`, `phi = 30`):

```

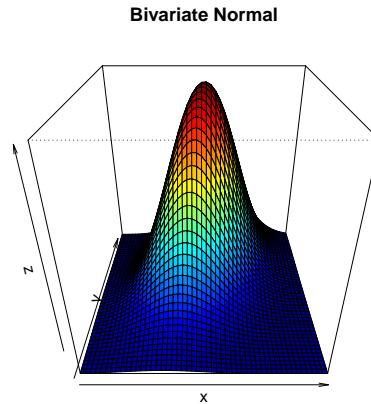
x <- seq(-3, 3, length = 50)
y <- seq(-3, 3, length = 50)
z <- outer(x, y, function(a, b) ...)

...

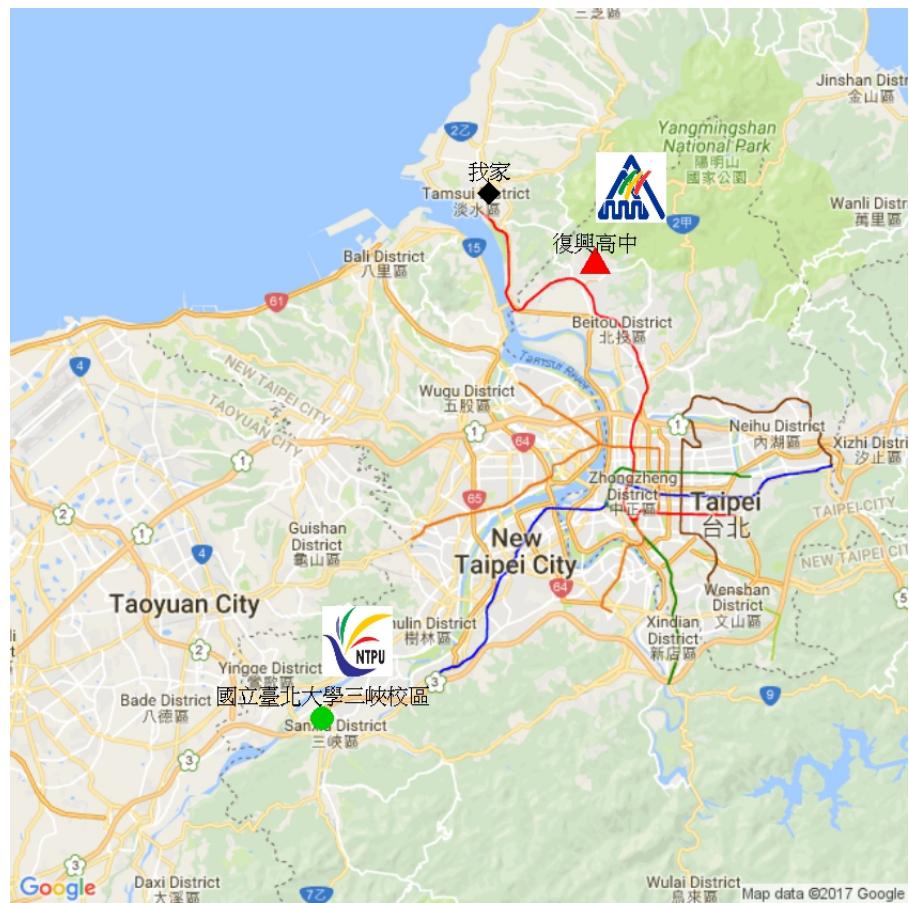
```



(d) 同上小題，以 `tim.colors {fields}` 為色階 (取 100 色)，畫出此二元常態聯合機率密度函數圖如下 (提示: `?persp`):

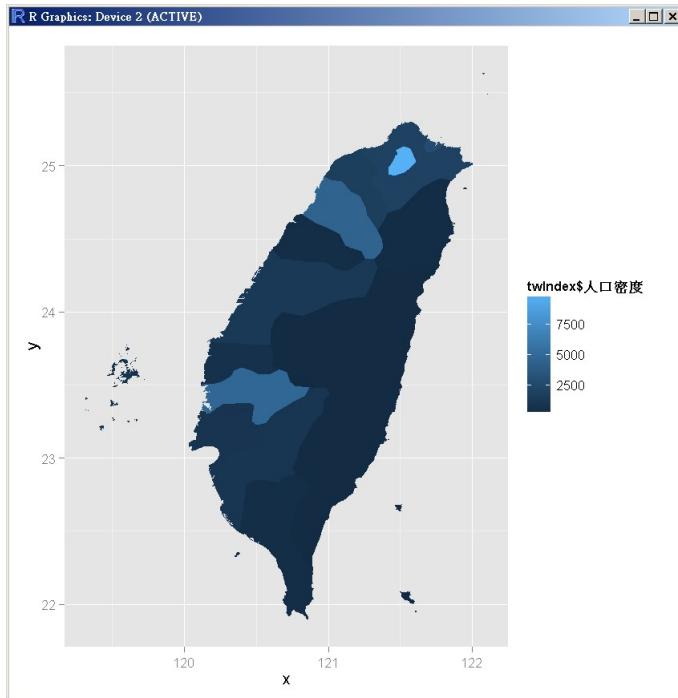


3.64 (地圖練習) 於台灣地圖上標記「國立臺北大學三峽校區」、「我唸的高中」及「我家」。(需同時標記符號及文字。我家限於台灣地區。將「國立臺北大學三峽校區」和「高中」的校徽貼在地圖上。高中校徽自己找喲!) (提示: `TextOnStaticMap`, `rasterImage`, `LatLon2XY.centered`)



3.65 (Choropleth Maps 練習) 資料來源: 中華民國統計資訊網 <http://statdb.dgbas.gov.tw>。

資料檔: stat.txt 台灣各縣市人口密度 : (人/平方公里) 參考: How to Make Choropleth Maps with R (<https://yaojenkuo.github.io/choroplethMap.html>) , 畫出下圖。



3.66 (Choropleth Maps 練習) state.x77 是 1977 年美國人口普查局針對全美 50 州發佈的一份調查紀錄。請利用 ggplot2 套件畫出 Population、Income、Murder 及 Illiteracy 的 Choropleth Maps。(自行選擇合適的色階)

```
> head(state.x77)
      Population Income Illiteracy Life_Exp Murder HS_Grad Frost Area
Alabama       3615    3624      2.1   69.05  15.1  41.3    20 50708
Alaska        365     6315      1.5   69.31  11.3  66.7   152 566432
Arizona       2212    4530      1.8   70.55   7.8  58.1    15 113417
Arkansas      2110    3378      1.9   70.66  10.1  39.9    65 51945
California    21198   5114      1.1   71.71  10.3  62.6    20 156361
Colorado      2541    4884      0.7   72.06   6.8  63.9   166 103766
> ?state.x77
```

3.67 統計資料之面量圖

- 資料來源: 新北市統計資料庫
- 資料集: 2018 年新北市各行政區之土地面積 (平方公里) 及人口密度資料
- 網址: http://pxweb.bas.ntpc.gov.tw/pxweb/dialog/statfile9_n.asp

請於上述網址下載，新北市統計資料庫: 二、人口: 現住戶數、人口密度及性比例)
「2018 新北市各行政區之土地面積 (平方公里) 及人口密度資料」。繪製新北市各行

政區土地面積 (平方公里) 及人口密度資料之面量圖。

3.68 地圖標記

- 資料來源: 7-11 便利商店
- 資料集: 7-11 新北市三峽區門市查詢-統一超商門市
- 網址: <http://www.i-write.idv.tw/life/info/7-11/711-304.html>
- 資料檔: SanShia7-11.csv

讀取資料檔 SanShia7-11.csv 中的地址資料，將之標記在 google 地圖上 (roadmap 型式)。

4 繪圖: ggplot2

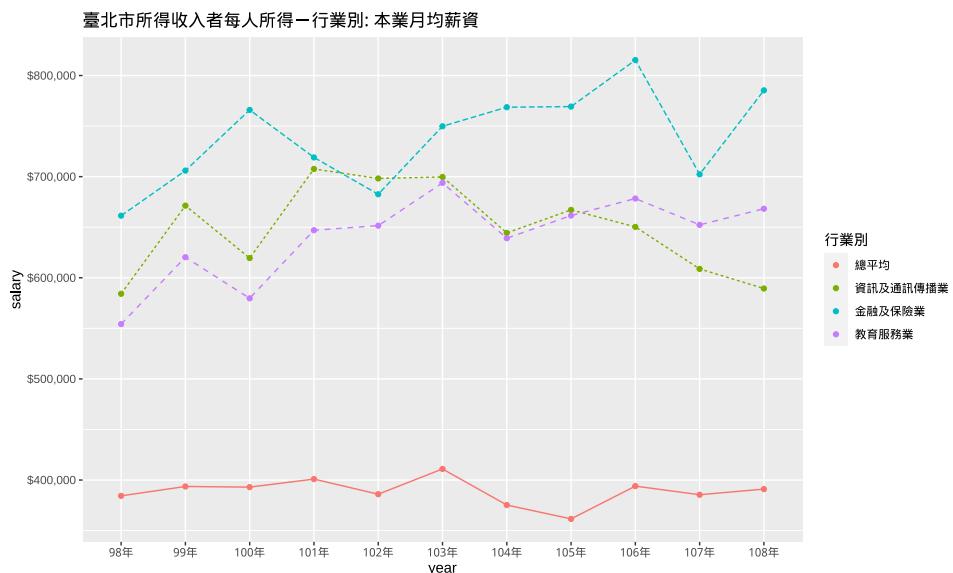
4.1 t 分佈在 wikipedia 中的介紹如下列網頁: https://en.wikipedia.org/wiki/Student%27s_t-distribution。以 ggplot2 套件，畫出 t 分佈在自由度為 1 及自由度為 5 的 (a) 機率密度函數圖，(b) 累積機率分佈函數圖，(c) 分位數函數圖及 (d) 隨機抽樣 ($n = 100$) 直方圖。(要求: (1) 前三個圖上各有兩條不同自由度之函數曲線 (以不同顏色表示)，直方圖則為重疊 (以不同顏色表示)。(2) 需加註: 標題， x 及 y 標號及 legend (以不同顏色表示相對應的自由度)。(3) 4 張圖一頁: 2 by 2)。

4.2 臺北市所得收入者每人所得，本業薪資 (依行業別)

- 資料敘述: 臺北市所得收入者每人所得 - 行業別-年依項目，年別與行業
- 資料來源: 台北市家庭收支資料庫查詢系統
- 資料網址: https://statdb.dbas.gov.taipei/pxweb2007-tp/dialog/statfile9_FI.asp
- 資料檔: fi00123y2a2021523444471.xlsx

讀取資料，並以 ‘ggplot2’ 畫出下圖。提示:

- `read_excel {readxl}` 會使用到的參數 (引數) 為 `range` 和 `na`。
- 檢查資料之結構，及每個欄位之資料類別是否正確。
- 讀入之資料，第一欄為年份 (命名 `year`)，並修正為有順序之因子。
- 選取繪圖所需之欄位資料，造出一個為 tidy 型式的資料框 (使用 `stack {utils}` 或 `melt {reshape2}`)，並重新對各欄位命名。
- 圖形中，有主標題，縱軸 tick 之標記有 \$ 之符號。

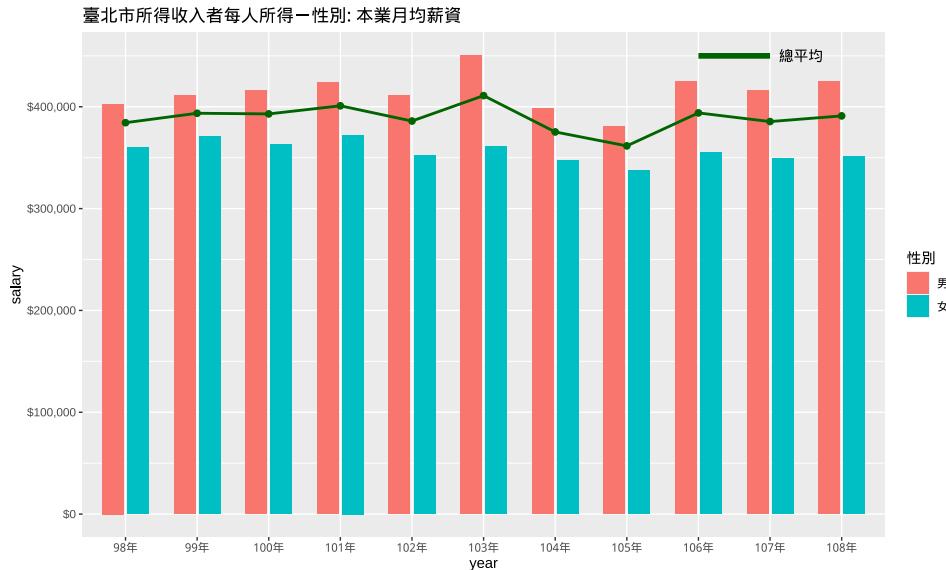


4.3 臺北市所得收入者每人所得，本業薪資 (依性別)

- 資料敘述: 臺北市所得收入者每人所得 - 性別-年依項目，年別與性別

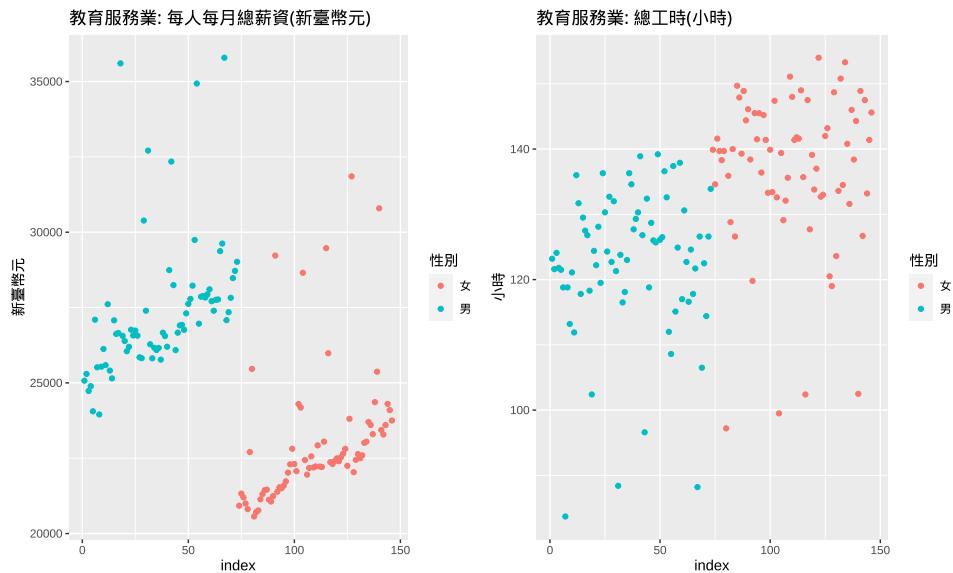
- 資料來源: 台北市家庭收支資料庫查詢系統
- 資料網址: https://statdb.dbas.gov.taipei/pxweb2007-tp/dialog/statfile9_FI.asp
- 資料檔: fi00128y2a2021523591831.xlsx

讀取資料，並以 ggplot2 畫出下圖。提示: geom_bar，geom_line，geom_segment，geom_text。



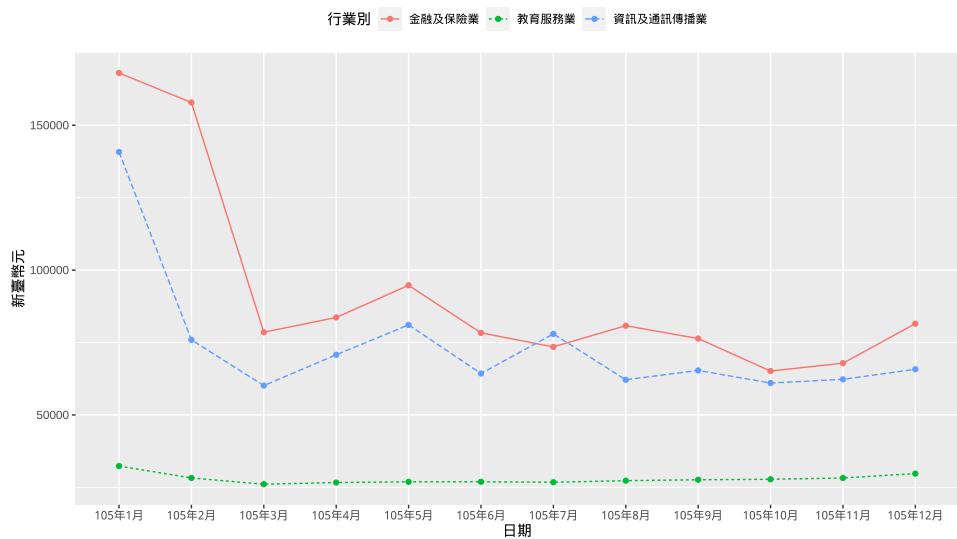
4.4 資料檔 SalaryGov.xlsx 是由行政院主計總處薪情平臺 <https://earnings.dgbas.gov.tw> 下載整理而得，包含 101 年 8 月至 107 年 8 月期間，全國不同行業別 (industry)，男女性 (gender) 的「每人每月總薪資 (新臺幣元)」(SalaryMonth 工作表) 及「總工時 (小時)」(TotalWorkHour 工作表)。

- (a) 讀取資料中的兩個工作表 (分別命名為 SalaryMonth 及 TotalWorkHour) 並分別印出前二個行業別 (礦業及土石採取業、製造業) · 男女性的統計值，前後各 5 筆紀錄。(前 5 筆紀錄: 101 年 8 月 ~101 年 12 月; 後 5 筆紀錄: 107 年 4 月 ~107 年 8 月) (提示: 讀檔分兩次讀，第一次讀取欄位名，做整理，第二次讀取資料，最後合併兩者。)
- (b) (索引圖) 若某人對「教育服務業」有興趣，請以 ‘ggplot2‘套件畫出「每人每月總薪資 (新臺幣元)」及「總工時 (小時)」的索引圖。(要求: 一頁兩張圖。圖上需以不同顏色標記性別。需有標題) (提示: 各別對於‘SalaryMonth‘及‘TotalWorkHour‘，先造出一個適合 ‘ggplot2‘繪製圖形的 tidy data 型式的資料框，欄位名為「日期」、「教育服務業」及「性別」。)

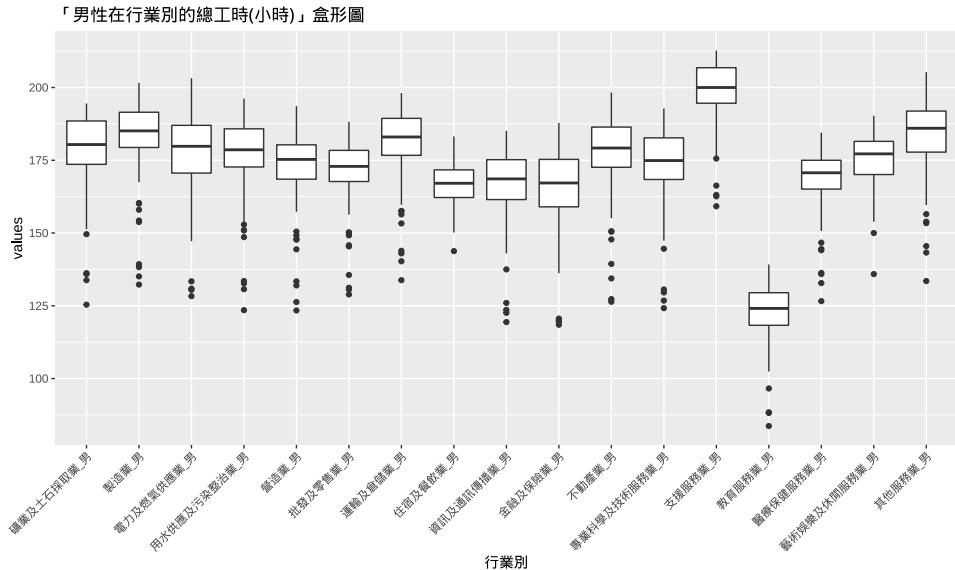


- (c) (折線圖) 以 ggplot2 套件畫出 105 年間，男性在「資訊及通訊傳播業」、「金融及保險業」及「教育服務業」的「每人每月總薪資 (新臺幣元)」時間序列圖。(提示: 對於 SalaryMonth，造出一個適合 ggolpt2 繪製圖形的 tidy data 型式的資料框，欄位名為「日期」、「行業別」。) (程式提示: grep("105", SalaryMonth\$ 日期), factor(SalaryMonth[id,]\$ 日期, levels= SalaryMonth[id,]\$ 日期, ordered = T))

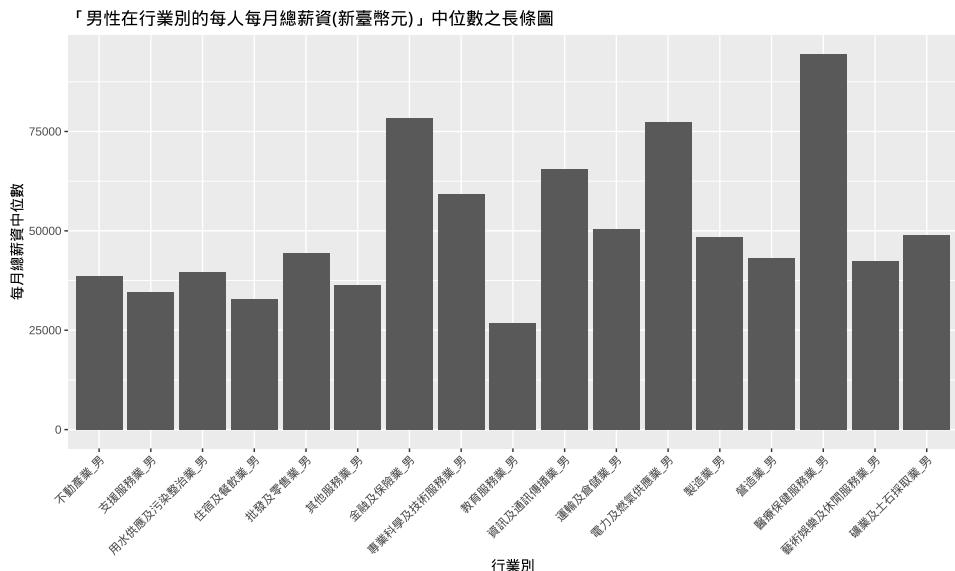
105年，每人每月總薪資(新臺幣元): 三行業之時間序列圖



- (d) (side-by-side 盒形圖) 以 ggplot2 套件畫出男性在各行業別中，「總工時 (小時)」的盒形圖。



(e) (長條圖) 以 ggplot2 套件畫出男性在各行業中，「每人每月總薪資 (新臺幣元)」中位數的長條圖。



4.5 (簡易資料處理、以 ggplot2 做基礎繪圖) 「‘diabetes.csv’」為一記錄糖尿病的資料集，其變數解釋如下：

- 編號 (id)
- 性別 (sex) : ‘1’= 女生, ‘2’= 男生
- 年齡 (age)
- 教育程度 (edu): ‘1’= 不識字, ‘2’= 小學, ‘3’= 國 (初) 中, ‘4’= 高中 (職), ‘5’= 大專以上
- 糖化血紅素 (a1c)
- 體重 (wt)
- 膽固醇 (ldl)
- 收縮壓 (sbp)

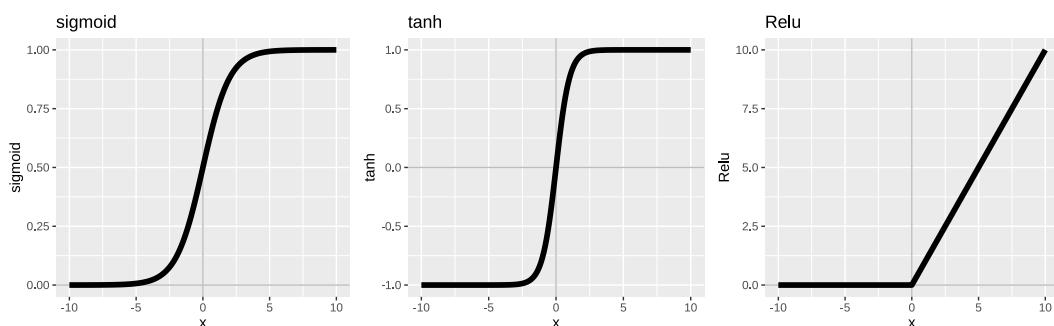
- 舒張壓 (dbp)
 - 嚴重度 (a1cgp) : '1' = 輕, '2' = 重
- (a) 讀入資料集「diabetes.csv」(命名為 diabetes) · 使用 xtable 套件 · 印出資料前後各 5 筆資料及其結構。(指令提示: xtable)
- (b) (資料處理) 判斷變數 sex, edu, a1cgp 是否為 factor 類別? 若不是, 請轉換成 factor 類別變數, 其中 edu 為有順序的 factor 類別變數。(要求: 需將上述三變數中的類別編碼轉回文字 · 例如: 性別 (sex) : '1' 轉為女生, '2' 轉為男生)。更改完後再印出資料結構。(指令提示: sapply · factor)
- (c) (資料處理) 印出哪些編號 (id) 的紀錄具有遺失值 (NA)? 刪除具有 NA 的紀錄 · 並將剩餘之完整資料另存成一 R 資料框 (命名為 diabetes.complete) · 其資料維度為何? 印出前 10 筆紀錄。(指令提示: apply · is.na)
- (d) (索引圖) 繪出 diabetes.complete 中連續變數的索引圖 (id 除外)。(要求: (1) 請六張圖一頁 (3 by 2); (2) 需加標題名 ("索引圖") 及座標名 (x 軸名為 'index', y 軸名為變數名); (3) 圖上點之顏色以性別 ('sex') 的類別為依據; (4) 圖上點之符號型狀以嚴重度 (a1cgp) 的類別為依據。)(指令提示: lapply · sapply · is.numeric · marrangeGrob)
- (e) (盒形圖) 各依照變數 sex, edu 及 a1cgp 之類別 · 繪出 diabetes.complete 中 · 膽固醇 (ldl) 之盒形圖。(要求: (1) 請三張圖並列一頁; (2) 需加標題及座標名)
- (f) (散佈圖) 繪出 diabetes.complete 中 · 兩個並列的散佈圖 (x vs y): (1) 體重 (wt) vs 膽固醇 (ldl); (2) 收縮壓 (sbp) vs 舒張壓 (dbp)。圖上點之顏色以性別 (sex) 的類別為依據; (4) 圖上點之符號型狀以嚴重度 (a1cgp) 的類別為依據。

4.6 (畫激勵函數) 深度學習 (deep learning) 領域中常使用的激勵函數 (activation function) (https://en.wikipedia.org/wiki/Activation_function) 有以下三種 · 請用 ggplot2 套件畫出它們的圖形。

$$\text{Sigmoid or logistic function (sigmoid)} : f(x) = \frac{1}{1 + e^{-x}}$$

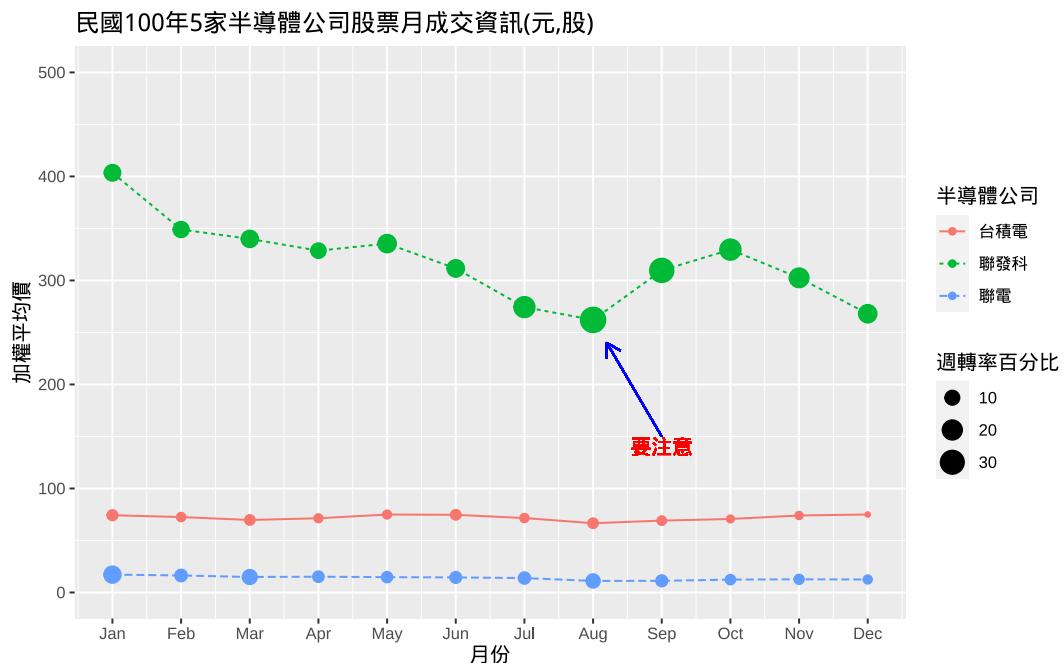
$$\text{Hyperbolic tangent function (tanh)} : f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\text{Rectified linear unit function (Relu)} : f(x) = xI(x \geq 0)$$



4.7 有民國 100 年 5 家半導體公司股票月成交資訊紀錄於資料檔 (stock-data.txt) 中。

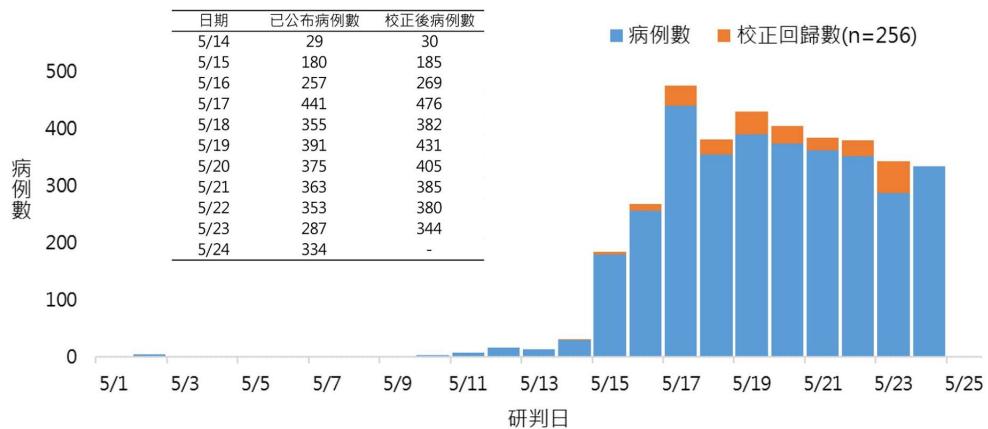
- 請讀入此資料並使用 ‘xtable‘套件，印出資料前 5 筆資料及其結構。
- 請將資料中的「成交筆數」、「成交金額」及「成交股數」轉成數值型變數後，印出資料前 5 筆資料及其結構。
- 請繪出以下三家公司之折線圖。(要求: (1) x 軸標記為月份英文簡記。(2) y 軸的範圍為 0~500)



4.8 (COVID-19 本土確定病例校正回歸情形)

下圖是台灣中央流行疫情指揮中心發佈的「5/1-5/23 COVID-19 本土確定病例校正回歸情形」長條圖，相關資訊可由衛生福利部疾病管制署 <https://www.cdc.gov.tw> 查詢得知。

5/1-5/23 COVID-19 本土確定病例校正回歸情形

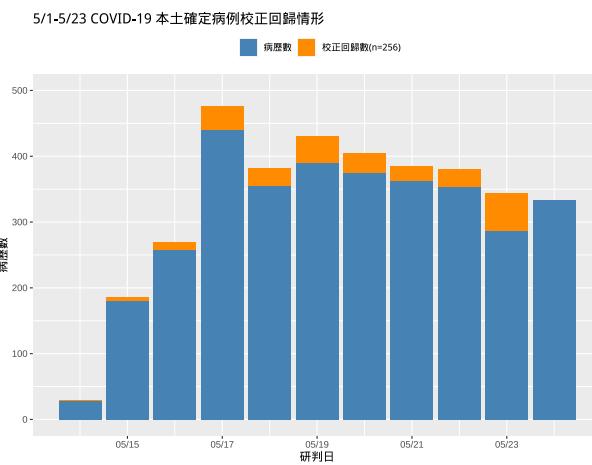


- **回歸病例定義**：個案採檢通報後，因檢體運送、實驗室量能、無法即時登打檢驗結果等因素，未於2日內完成檢驗結果報告之個案。
- **研判基準**：確診個案之採檢日至確診日(陽性檢驗結果通知)超過2日以上。

中央流行疫情指揮中心

2021/05/24 更新版

- (a) (tidy data) 請依據圖中之資料表格，造出一個便於使用 `ggplot2` 繪圖的資料框 (命名 `covid19.data.tidy`)。
- (b) 承上題，以 `ggplot2` 套件畫出下圖。(提示: `scale_fill_manual` , `scale_x_date`)

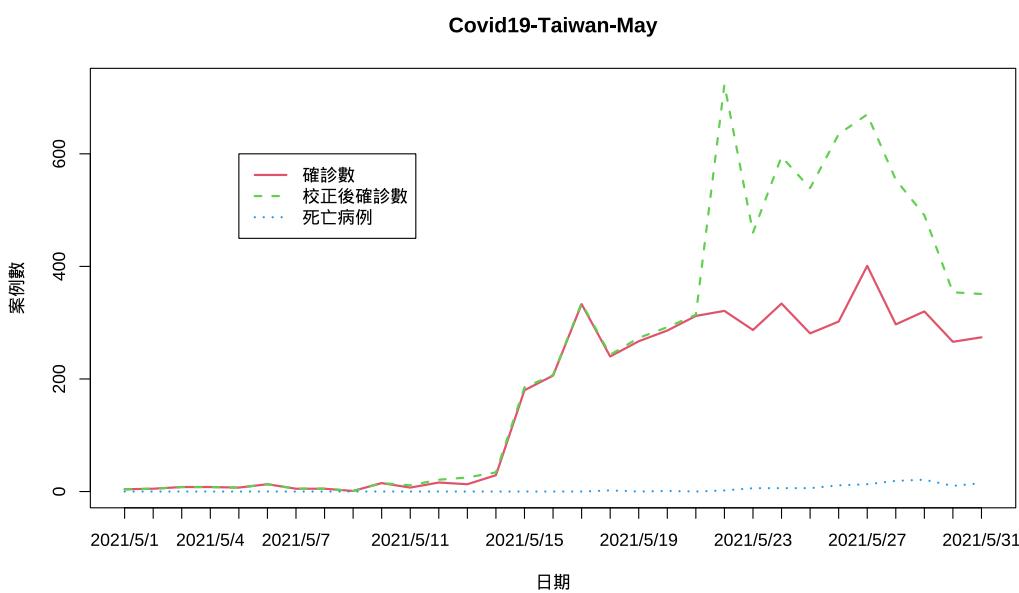


- (c) 承上題，以套件 `ggthemes`，畫出三個不同佈景主題之圖形。

4.9 (台灣 COVID-19 五月份資料) 下圖為「聯合新聞網: 追台灣最新疫情看即時數據圖表」https://topic.udn.com/event/COVID19_Taiwan」所發佈的台灣 5/1-5/31 COVID-19 本土確診病例及校正回歸數之長條圖和死亡病例的折線圖，同時有標註四個警示事件。

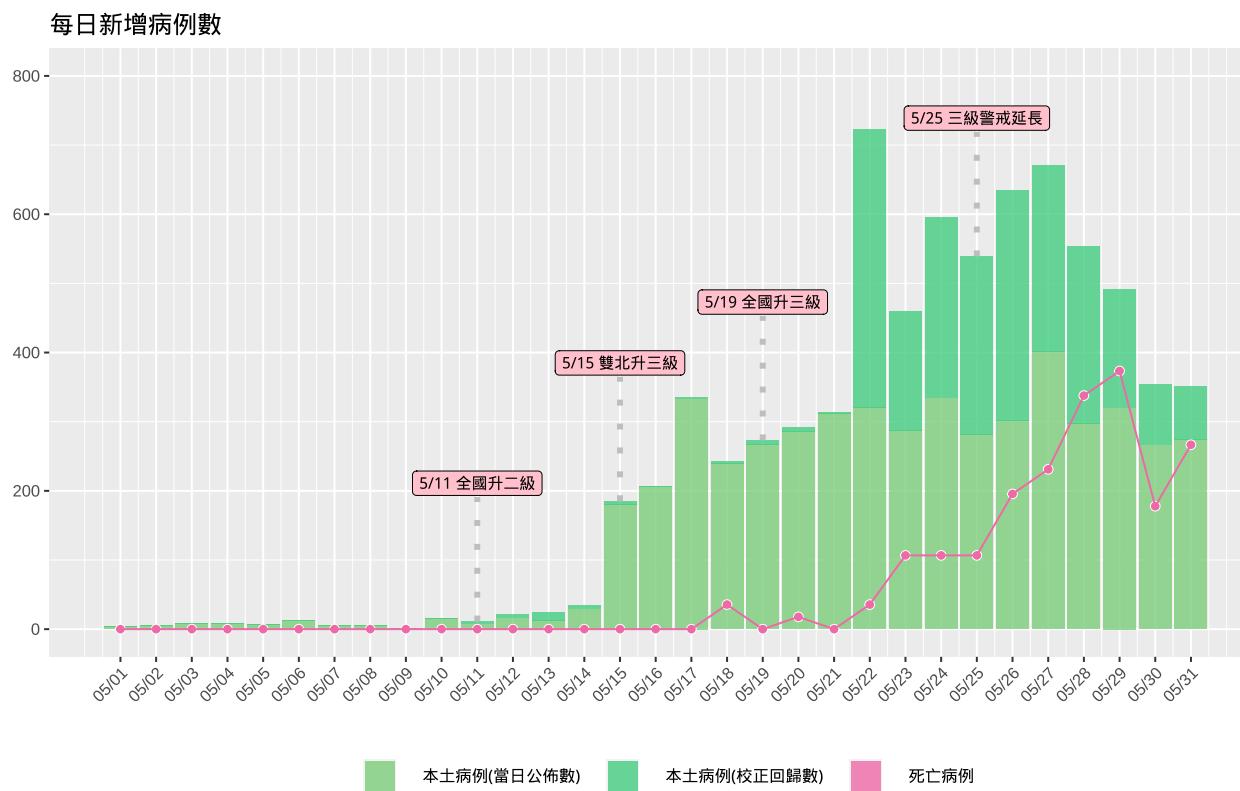


- (a) 資料檔 Covid19-Taiwan-May.csv 來自於「財團法人國家實驗研究院國家高速網路與計算中心:COVID-19 疫情紀錄表單」https://covid-19.nchc.org.tw/dt_005-covidTable_taiwan.php，收集五月份台灣 COVID-19 的確診病例、校正回歸數及死亡病例的數據。將此資料入 R(命名 Covid19.Taiwan.May.df)並印出。
- (b) 畫出下列時間序列圖。



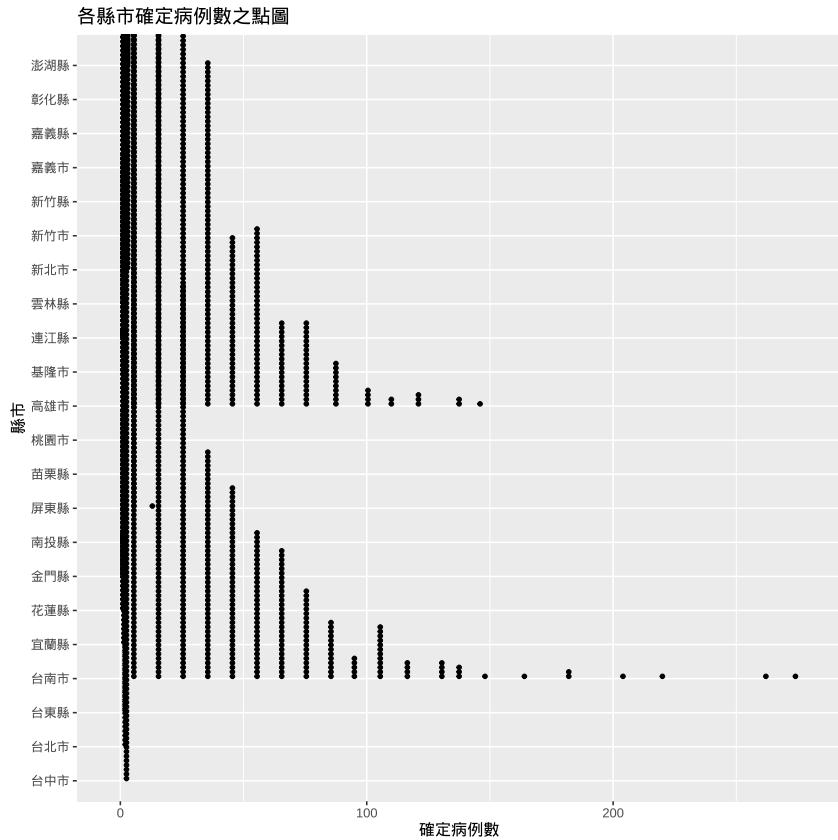
- “
- (c) 若要繪製出圖 (1)，請依據 1.1 題讀入之資料，造出一個便於使用 ggplot2 繪圖的資料框 (命名 Covid19.Taiwan.May.tidy)。
- (d) 承上題，以 ggplot2 套件畫出圖 (1)。(繪製出來圖要儘量與圖 (1) 相似，例如：四次事件標註；右側的縱座標。)(提示：library(ggnewscale)，geom_bar，

```
scale_x_date, geom_point, geom_line, scale_fill_manual, theme, scale_y_continuous,
geom_label, geom_segment, labs)
```

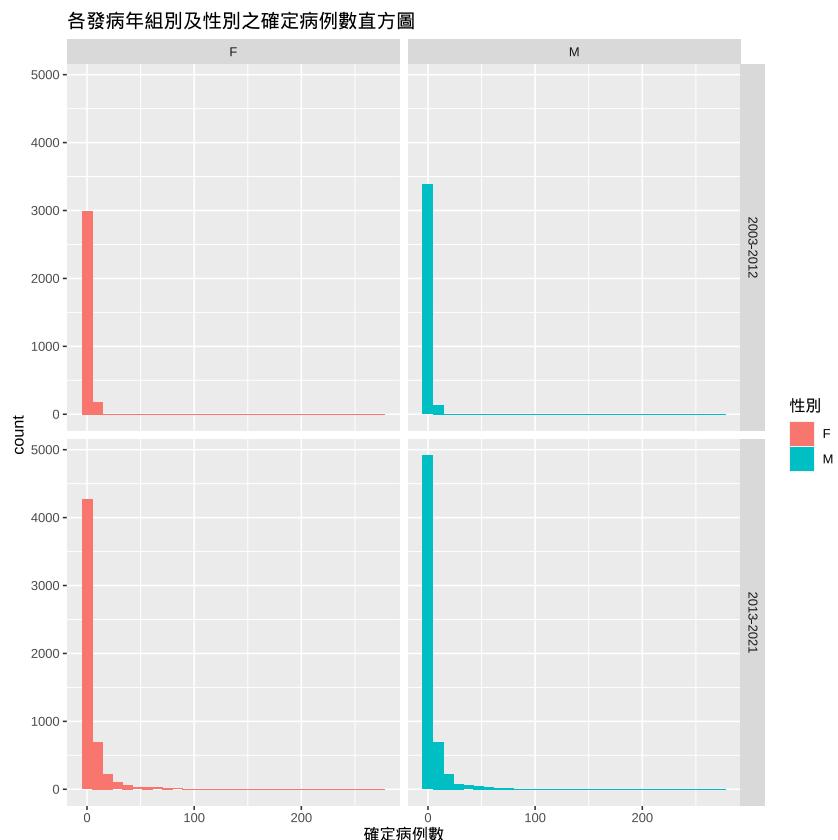


4.10 登革熱病媒蚊調查資料

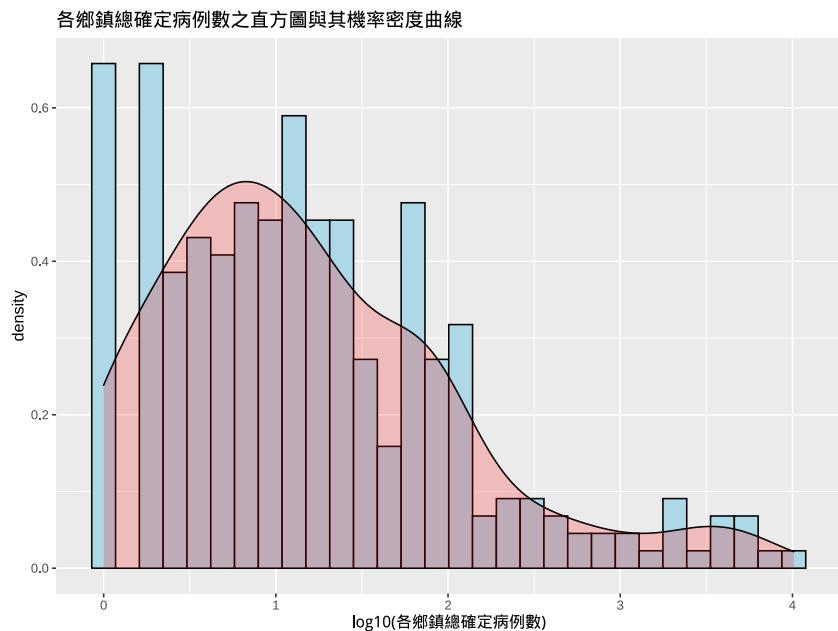
- 資料名稱: 登革熱病媒蚊調查資訊
 - 資料來源: 衛生福利部疾病管制署 <https://data.cdc.gov.tw/dataset/aagstable-dengue>
 - 資料檔: 見所附檔案Age_County_Gender_061.csv
 - 資料說明: 2003-2021 年間，各地區、各年齡層之登革熱病例數統計表。每一筆紀錄之欄位包含確定病名、發病年份、發病月份、縣市、鄉鎮、性別、是否為境外移入、年齡層及確定病例數。
- (a) 讀入資料集「Age_County_Gender_061.csv」(命名為 Dengue.fever)，使用 xtbl 套件，印出資料前後各 5 筆資料。
 - (b) 印出 Dengue.fever 之結構及摘要。
 - (c) 繪出本資料中，(a) 各‘發病年份’之‘確定病例數’的盒形圖及 (b) 各‘發病年份’之‘確定病例數’(在 \log_{10} 尺度下) 的盒形圖如下。
 - (d) 繪出本資料中，2003-2021 年間，各‘縣市’之‘確定病例數’之點圖如下。



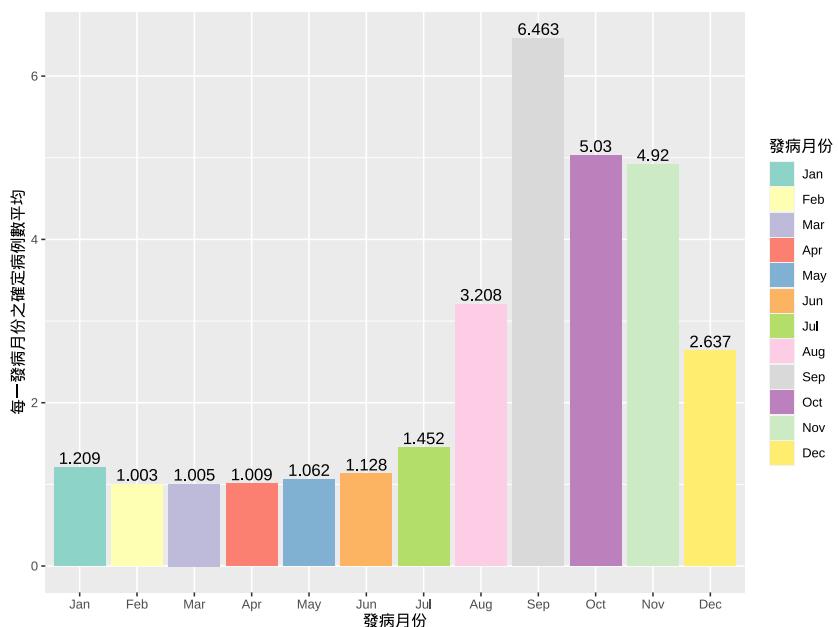
- (e) 將發病年份依 2003-2012 及 2013-2021 分成兩組 ('發病年組別')。依此兩組，畫出各 '性別' 之 '確定病例數' 直方圖。



\log_{10} 之後，繪出直方圖並加機率密度曲線。



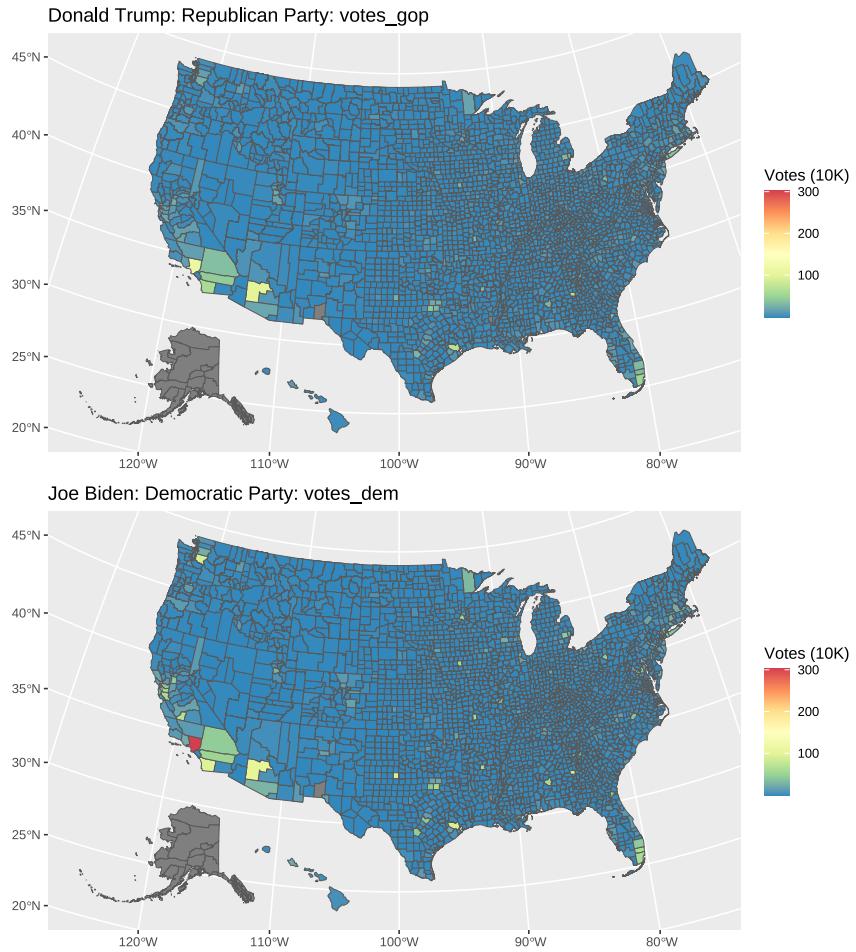
- (g) 繪出資料中，每一‘發病月份’之‘確定病例數’平均之長條圖如下圖。(要求：使用 RColorBrewer 之 Set3 色階填滿)



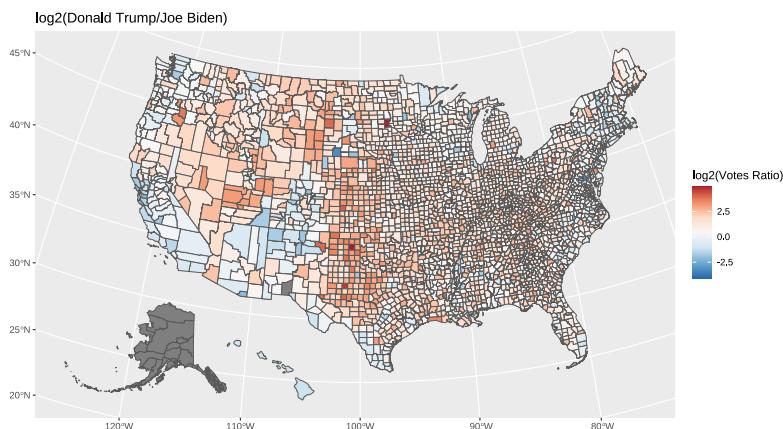
4.11 2020 美國大選，候選人各地區 (county) 得票數之面量圖

- 資料來源: https://github.com/tonmcg/US_County_Level_Election_Results_08-20
- 檔案名稱: 2020_US_County_Level_Presidential_Results.csv

- (a) 請畫出「2020 美國大選，Donald Trump 和 Joe Biden 兩位候選人在美國各區 (County) 的得票數」之面量圖。(要求：兩張圖使用的色階，尺度需一致。)



- (b) 在美國各區，將 Donald Trump 之得票數除以 Joe Biden 之得票數，再取以 2 為底之對數 (log)，當成兩候選人在美國各區得票領先或落後之指標，請在美國各區畫出此指標。(要求：色階需為雙向色階。)



4.12 臺中市 108 年各行政區之「人口密度 (人 / 平方公里)」面量圖

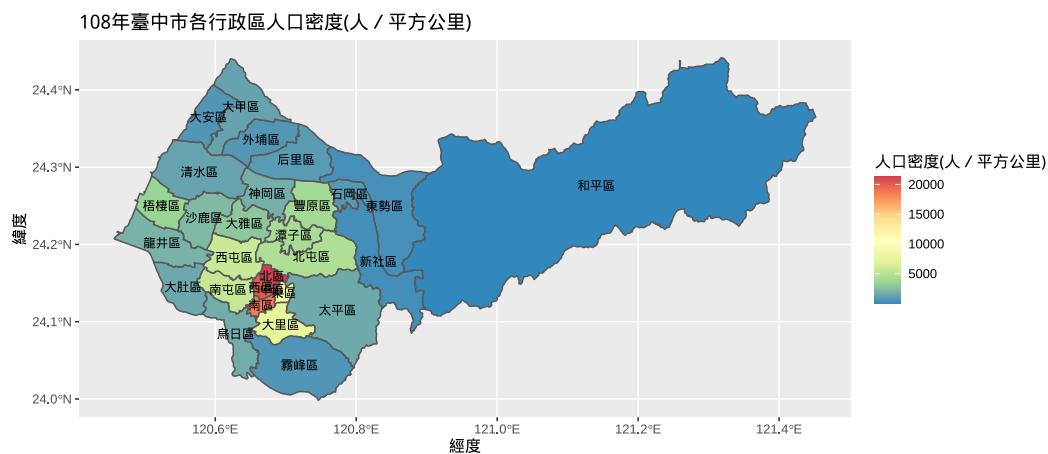
資料來源: 臺中市統計資料查詢平臺 (<https://govstat.taichung.gov.tw/DgbasWeb/>)

`index.aspx`)。

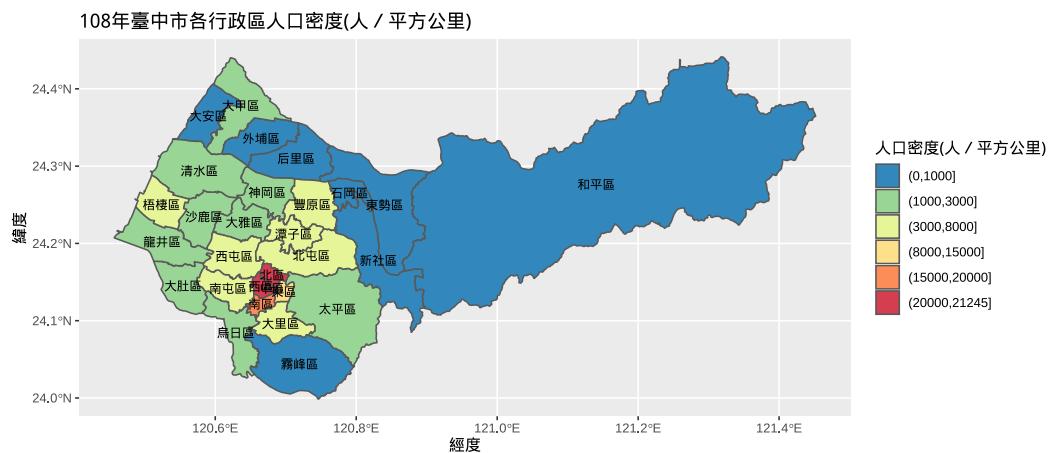
說明: 查詢方式為 . 資料表式-統計年報 (行政區資料) \Rightarrow 貳、人口 \Rightarrow 臺中市人口、密度及性比例 \Rightarrow 時間: 107-108; 地區: 全選; 臺中市人口、密度及性比例。

檔案名稱: `taichung-pop-query.xlsx`

- (a) 請畫出 108 年臺中市各行政區之「人口密度 (人 / 平方公里)」。(要求: 地圖來源需使用「Shape files」，各行政區需標上行政區名稱。顏色之色階為單向連續型色階)。(提示: `as.numeric(gsub(", ", "", "123,456"))`)



- (b) 將「人口密度 (人 / 平方公里)」以下列六個區間分組: 「 $0 - 1000, 1001 - 3000, 3001 - 8000, 8001 - 15000, 15001 - 20000, 20001 - \infty$ 」，選取合適之連續型色階 (6 個顏色對應至前述之分組)，重新繪制上述地圖。



4.13 (Google Map 標示位置) 請在台灣地圖上以「圖片」、「標題」及「經緯度」標示出下列地點 (資料來源: 自行收集):

- (a) 標題: 北大三峽校區 (圖片: 學校 logo)。
- (b) 標題: 我的家 (圖片: 自己的大頭照)。
- (c) 標題: 總統府 (圖片: 總統府)。
- (d) 標題: 最想去玩的兩個景點 (圖片: 景點照)。

5 微積分、線性代數

5.1 有三個矩陣如下，計算 (a) AB 。 (b) $2A + 3C^T$ 的矩陣運算結果。

$$A = \begin{bmatrix} 2 & 4 & -1 \\ 5 & 8 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & -5 & 1 & 4 \\ 4 & 2 & 0 & 3 \\ -3 & 1 & 2 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & -1 \\ 8 & -3 \\ -6 & 2 \end{bmatrix}.$$

5.2 寫一 R 函式 (命名為 `my.inverse`)，使其執行時會要求使用者輸入一 3×3 矩陣 (`scan`)，在計算其反矩陣後，回傳原始輸入矩陣及其反矩陣。(提示：`solve`)。請利用以下矩陣做測試：

$$\begin{bmatrix} 3 & 5 & -1 \\ 2 & -1 & 3 \\ 4 & 2 & 3 \end{bmatrix}$$

5.3 設 $A = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 3 \end{bmatrix}$, $C = \begin{bmatrix} 3 & 0 & 4 \\ 1 & 1 & 7 \end{bmatrix}$ 。求下列各題中的矩陣 X 。

- (a) $A - B + 2X = C + 3A + X$
- (b) $3X + A + B + C = A - 2B + 4C + X$

5.4 設 $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 & 8 \\ 6 & 2 & 5 \end{bmatrix}$ ，求 $(A + B)^T$ 和 $A^T + B^T$ 。

5.5 設 $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 2 & 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 & 1 \\ 3 & 2 & 4 \\ 3 & 5 & 1 \end{bmatrix}$ ，(1) 試求 A^{-1} 。 (2) 若 $AX = B$ ，求 X 。
(3) 若 $XA = B$ ，求 X 。

5.6 一個 2×2 矩陣 $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ 的反矩陣公式為

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

試寫一 R 函式，(a) 求一 2×2 矩陣之反矩陣，以 $A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$ 為例。(b) 與 R 內建函式 `solve` 相比較。

程式提示:

```
A <- ...
A.inverse <- function(A){
  ...
}
# (a)
A.inverse(A)
...

# (b)
solve ...
```

5.7 有 $A \cdot B$ 兩矩陣如下:

$$A = \begin{bmatrix} 3 & -2 \\ 1 & 4 \\ 2 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 2 & 1 \end{bmatrix}.$$

試求 $A^T + 2B$ 、 AB 、 BA 。

5.8 令 $\mathbf{u} = (2, 4, 4)$, 試求在 \mathbf{u} 方向的單位向量。(若 $\mathbf{x} = (x_1, x_2, \dots, x_p)$ 為一向量, 則在其方向的單位向量為 $\frac{1}{\|\mathbf{x}\|} \mathbf{x}$, 其中 $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$) (提示: $\mathbf{u} <- \mathbf{c}(2, 4, 4)$)

5.9 兩向量 $\vec{u} = (u_1, u_2, \dots, u_p)$ 及 $\vec{v} = (v_1, v_2, \dots, v_p)$ 的 cosine 夾角 θ , 可用下列式子表示:

$$\cos \theta = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|},$$

其中 $\vec{u} \cdot \vec{v} = \sum_{i=1}^p u_i v_i$, $\|\vec{u}\| = \sqrt{\sum_{i=1}^p u_i^2}$ 和 $\|\vec{v}\| = \sqrt{\sum_{i=1}^p v_i^2}$ 。現若有兩向量 $\vec{u} = (3, 4, 0)$ 和 $\vec{v} = (4, 4, 2)$, 試計算此兩向量之的 cosine 夾角 ($\cos \theta$)。

5.10 空間上有兩平面 $I_1 : a_1x + b_1y + c_1z + d_1 = 0$ 及 $I_2 : a_2x + b_2y + c_2z + d_2 = 0$, 其法線 (normal) 分別為 $\vec{n}_1 = (a_1, b_1, c_1)$ 及 $\vec{n}_2 = (a_2, b_2, c_2)$, 若 $\vec{n}_1 = k\vec{n}_2$, 亦即 $\frac{a_1}{a_2} = \frac{b_1}{b_2} = \frac{c_1}{c_2} = k$, 則我們稱此兩平面平行。而此兩平行的平面, 其距離公式為:

$$D = \frac{|a_2x_0 + b_2y_0 + c_2z_0 + d_2|}{\sqrt{a_2^2 + b_2^2 + c_2^2}},$$

其中 $P_0 = (x_0, y_0, z_0)$ 為平面 I_1 上之一點。若兩平面分別為 $I_1 : 2x + 3y + 4z - 3 = 0$ 及 $I_2 : -4x - 6y - 8z + 8 = 0$, $P_0 = (0, 0, 3/4)$ 為平面 I_1 上之一點, 試回答下列各問題:

- (a) 將 \vec{n}_1 、 \vec{n}_2 及 P_0 表示為 R 向量類別的物件。
- (b) 利用 \vec{n}_1 , \vec{n}_2 , 說明平面 I_1 和 I_2 平行。
- (c) 利用 \vec{n}_1 、 \vec{n}_2 及 P_0 計算此兩平行的距離。

5.11 有一數學函數定義在整個實數線上，如下：

$$f(x) = \begin{cases} -x, & x < 0 \\ x^2, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

若給定 x 值為 $-2, -1.5, 0, 0.7, 1, 3.6$ ，試寫一 R 函式 fn 算出函數值 $f(x)$ 如下：

```
> x
[1] -2.0 -1.5  0.0  0.7  1.0  3.6
> fn(x)
[1] 2.00  -    -    -    -    -
```

5.12 一連續函數 f 若滿足 $f(-x) = f(x)$ ，則稱 f 為偶函數 (even)。若 f 滿足 $f(-x) = -f(x)$ ，則稱 f 為奇函數 (odd)。試寫一 R 函式 (命名為 check_even_odd)，以數值的方法，判別以下三個函數是偶函數或是奇函數。

$$f_1(x) = -\frac{8}{x^2 - 4}, \quad -2 < x < 2.$$

$$f_2(x) = \frac{4x}{\sqrt{x^2 + 1}}, \quad -\sqrt{3} < x < \sqrt{3}.$$

$$f_3(x) = x^3(1 + x^4)^3, \quad -1 < x < 1.$$

```

f1 <- function(x){
  ...
}

f2 <- function(x){
  ...
}

f3 <- function(x){
  ...
}

check.even.odd <- function(f, x){
  ...

  cat(" => 此函數為偶函數 。\n")
  ...
}

x1 <- seq(-2, 2, length.out=100)
check_even_odd(f1, x1)
=> 此函數為偶函數 。

x2 <- seq(0, sqrt(3), length.out=50)
check_even_odd(f2, x2)
.....

x3 <- ...
.....

```

5.13 Compute the values of $f(x) = \frac{\sqrt{x^2 + 100} - 10}{x^2}$ when x near 0 °. (印出下表)

x	f(x)
1	...
0.5	...
0.1	...
0.01	...
0.0005	...
0.0001	...
0.00001	...
0.000001	...
-0.000001	...
-0.00001	...
-0.0001	...
-0.0005	...
-0.01	...
-0.1	...
-0.5	...
-1	...

5.14 某一連續函數 f 在 x_0 的導數 (derivative) 定義為

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

當 h 很小且 $h > 0$ 時, 以數值方式計算 $f'(x_0)$ 的一種方式是”The forward-difference formula”:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi), \quad \text{for some } \xi(x) \in (x_0, x_0 + h).$$

其中就是利用差商 (difference quotient, DQ) 來逼近 $f'(x_0)$:

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h}.$$

以差商來逼近 $f'(x)$ 的誤差上界是 $M|h|/2$, 其中 M 是 $|f''(x)|, x \in (x_0, x_0 + h)$ 的上界。請以差商逼近 $f(x) = \ln x$ 在 $x_0 = 1.8$ 的導數, 印出下表。(提示: 誤差上界為 $|h|/(2(1.8)^2)$)

h	f(1.8+h)	DQ	ErrorBound
0.1	0.64185389	0.5406722	0.0154321
0.05
0.01
0.001			

程式提示:

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

DQ <- function(f, x0, h){
  ...
  # compute error bound here
  ...
}

h <- c(0.1, 0.05, 0.01, 0.0001)
DQ(f, x0=1.8, h) #以下印出表格
h      f(1.8+h)      DQ      ErrorBound
=====
0.1    0.64185389    0.5406722   0.0154321
0.05   ...           ...       ...
0.01   ...           ...       ...
0.001
```

5.15 對一個在閉區間 $[a, b]$ 有定義的實數函數 f ，定義其黎曼和 (Riemann sum) 為以下式子：

$$S_P = \sum_{i=1}^n f(c_i)\Delta x_i, \quad \text{其中}$$

- $P = \{x_0 = a, x_1, \dots, x_{n-1}, x_n = b\}$ 為 $[a, b]$ 之分割 (partition)，
- $\Delta x_i = x_i - x_{i-1}$, $i = 1, \dots, n$,
- $c_i \in [x_{i-1}, x_i]$, $i = 1, \dots, n$, 常用的三種不同取法如下：
 - (i) 若 $c_i = x_{i-1}$ ，則 S_P 稱為下和 (lower sum)。
 - (ii) 若 $c_i = x_i$ ，則 S_P 稱為上和 (upper sum)。
 - (iii) 若 $c_i = (x_{i-1} + x_i)/2$ ，則 S_P 稱為使用子區間中點之和 (sums using the midpoints of each subinterval)。

今給定一函數 $f(x) = x^2 - 1$ 定義在 $[0, 2]$ 上，將 $[0, 2]$ 等距分割成 $n = 200$ 個子區間，(亦即 $\Delta x_i = (b - a)/n$)。試寫一 R 函式 (命名 `RiemannSum` 如下)，計算三種黎曼和。

(提示：先產生數列 $\{x_0, x_1, \dots, x_n\}$ ，再計算不同取法的 c_i 及 $f(c_i)$)

```
RiemannSum <- function(f, a, b, n){
  ...
}

> RiemannSum(f = myf, a = 0, b = 2, n = 200)
$lower.sum
[1] 0.6467

$upper.sum
[1] 0.6867

$sum.midpoints
[1] 0.66665
```

5.16 牛頓法求 $f(x) = 0$ 的解 (Newton's Method) 過程如下:

先猜測一初始值 x_0 為近似 $f(x) = 0$ 的根，再以初始值 x_0 代入下列迭代公式求得第一次近似根 x_1 ，如此一直重覆此過程而得到近似解:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{if } f'(x_n) \neq 0.$$

- (a) 若某一函數為 $f(x) = x^3 - x - 1$ ，其第一階導函數為 $f'(x) = 3x^2 - 1$ ，試寫兩個 R 函式 (各命名為 f 及 fp)，可計算其函數值 $f(x)$ 及其第一階導數值 $f'(x)$ 。
- (b) 呈上小題，試寫一 R 函式 (命名為 Newton_Method)，利用牛頓法求 $f(x) = 0$ 的正根 ($x_0 = 1$)。(至少迭代 5 次以上)

n	xn	f(xn)	f'(xn)	x(n+1)
0	1	-1	2	1.5
1	1.5	. . .		
. . .				

- (c) 若某一函數為 $f(x) = x^2 - 2 = 0$ ，求解 $f(x)$ 的正根，初始值為 $x_0 = 1$ 。(至少迭代 5 次以上)，並與 $\sqrt{2}$ 之結果相比較。

5.17 一個函數 f 在區間 $[a, b]$ 的定積分定義為: $\int_a^b f(x) dx = \lim_{\|P\| \rightarrow 0} \sum_{i=1}^n f(\bar{x}_i) \Delta x_i$, 其中

- P is a partition of the interval $[a, b]$ with n subintervals:
 $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$.
- Δx_i is the width of for the i th subinterval, $\Delta x_i = x_i - x_{i-1}$.

- \bar{x}_i is a sample point for the i th subinterval: $[x_{i-1}, x_i]$.
- $\|P\|$ denotes the length of the longest of the subintervals of the partition P .

利用上述定義計算 $\int_0^1 \sqrt{4 - x^2} dx$ 。令 $f(x) = \sqrt{4 - x^2}$ ，請依底下步驟做答。

- 請將區間 $[0, 1]$ 等分割為 10 等份，列印出子區間的端點值: $x_0, x_1, x_2 \dots, x_{10}$ 。
- 計算並列印出數列 $\Delta x_i = x_i - x_{i-1}$, $i = 1, \dots, 10$ 。
- 令 $\bar{x}_i = x_i$ ，計算並列印出數列 $f(\bar{x}_i)$, $i = 1, \dots, 10$ 。
- 計算 $\sum_{i=1}^{10} f(\bar{x}_i) \Delta x_i$ 的值。
- 將上述步驟 (a)~(d) 寫成一個函式 (命名為 `my.int`)，輸入為分割數 n (內定值為 10)，輸出為 $\sum_{i=1}^n f(\bar{x}_i) \Delta x_i$ 的值。
- 呈上題 (e)，求分割數 $n = 50, n = 100, n = 200, n = 2000, n = 5000$ 時的答案。
- 請利用 `integrate` 指令計算 $\int_0^1 \sqrt{4 - x^2} dx$ 。

5.18 一個函數 $f(x)$ 在 $[a, b]$ 之定積分可由合成的梯形法 (Composite Trapezoidal Rule) 來逼近。其公式為

$$\int_a^b f(x) dx \approx \frac{h}{2} \left[f(a) + 2 \sum_{j=1}^{n-1} f(x_j) + f(b) \right],$$

其中 $h = (b - a)/n$ ， $x_j = a + jh$, $j = 0, 1, \dots, n$, $a = x_0, b = x_n$. 試寫一個合成的梯形法的 R 函式，計算 $\int_0^2 x^2 \ln(x^2 + 1) dx$, $h = 0.25$ 之逼近值。(註：輸入為 a, b, h ，輸出為積分逼近值。)

5.19 R 軟體中計算積分之指令為 `integrate`，試計算 $\int_0^2 x^2 \ln(x^2 + 1) dx$ 。

5.20 假設 $\{x_0, x_1, \dots, x_n\}$ 是在某一區間 I 不同的 $(n + 1)$ 個數值點，且某一函數 f 在此區間是連續且可微的。以數值方式逼近此函數 f 在 x_0 的微分值 $f'(x_0)$ ，文獻上有以下兩個著名三點公式：

- 三點端點公式 (Three-Point Endpoint Formula):

$$f'(x_0) \approx \frac{1}{2h} [-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)].$$

- 三點中點公式 (Three-Point Midpoint Formula):

$$f'(x_0) \approx \frac{1}{2h} [f(x_0 + h) - f(x_0 - h)].$$

現假設有一函數 $f(x) = xe^x$ ，其部份資料點如下所示，試使用此資料逼近 $f'(2.0)$ 。

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

```

f <- function(x){
  ...
}

TPEF <- function(f, x0, h){
  ...
}

TPMF <- function(f, x0, h){
  ...
}

my.derivative <- function(...){
  ...
}

my.derivative(...)

## 執行my.derivative(...), 印出下表
The approximation of f'(2.0).
x0      h      TPEF      TPMF
=====
1.9    0.1    ...    ...
1.8    0.2    ...    ...
2.0   -0.1    ...    ...

```

5.21 一個連續函數的固定點 (fixed-point) 定義如下:

Let $g : R \rightarrow R$ be a continuous function. A *fixed point* of g is a number a such that $g(a) = a$. That is, a is a solution of the equation $f(x) = g(x) - x = 0$.

利用固定點法 (Fixed-point method) 求一個函數 $f(x) = 0$ 的根 $x = a$ 。其循環公式如下:

$$x_{n+1} = g(x_n),$$

其中 x_{n+1} 為 $f(x) = 0$ (即 $g(x) = x$) 之第 $n + 1$ 次根的近似值。期望當 n 夠大時， x_{n+1} 可以趨近 a 。利用固定點法求根的程式實作演算法如下：

Step 1: 輸入：初始值 (x_0)、函數 (g)、兩個演算法停止法則參數：(i) 容忍度 (tol) 要在一定的範圍內 (預設值： $|x_n - x_{n-1}| \leq tol$, $tol = 10^{-9}$); (ii) 達到迭代最大的次數 ($max.iter$) (預設值： $n = max.iter = 50$)，兩法則任一成立，演算法即停止。

Step 2: 循環計算 $x_{n+1} = g(x_n)$ ，並判別演算法是否需停止。(提示：`while`, `&&`, `cat`)

Step 3: 輸出：(1) 迭代次數及解的近似值。(2) 若演算法收斂，則印出 `Algorithm converged`，若演算法發散，則印出 `Algorithm failed to converge`。(定義演算法收斂為 $|x_n - x_{n-1}| \leq tol$ ，否則為發散)

利用固定點法和下列三種不同型式的 g 函數解 $f(x) = \log(x) - \exp(-x) = 0$ 之根，初始值為 $x_0 = 2$ ，容忍度為 $tol = 1e-06$ 。

- (a) $g_1(x) = \exp(\exp(-x)) = x$
- (b) $g_2(x) = x - \log(x) + \exp(-x) = x$
- (c) $g_3(x) = x + \log(x) - \exp(-x) = x$

執行結果示意畫面如下。

```

g1 <- function(x){
  exp(exp(-x))
}

g2 <- function(x){
  ...
}

g3 <- function(x){
  ...
}

fixedpoint <- function(g, x0, tol = 1e-9, max.iter = 50) {
  ...
}

> #1(a)
> fixedpoint(g1, 2, tol = 1e-06)
At iteration 1 value of x is: 1.144921
...
Algorithm converged
> #1(b)
> fixedpoint(g2, 2, tol = 1e-06)
At iteration 1 value of x is: 1.442188
...
Algorithm converged
> #1(c)
> fixedpoint(g3, 2, tol = 1e-06, max.iter = 20)
At iteration 1 value of x is: 2.557812
...
Algorithm failed to converge

```

5.22 紿定一矩陣 A 及一向量 \vec{b} 如下：

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 6 & -4 & 5 & -3 \\ 8 & -4 & 1 & 0 \\ 4 & -1 & 0 & 7 \end{bmatrix} \quad \text{and} \quad \vec{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}.$$

- (a) 於 R 中，輸入上述矩陣（命名為 A ）及向量（命名為 b ），並印出。
- (b) 請使用「R 指令」安裝可做「LU 分解」的 R 套件。（請自行 google 搜尋合適

的套件)。

- (c) 使用上述所安裝套件提供的指令，對 A 矩陣做 LU 分解，印出「L」、「U」及此兩矩陣之乘積「LU」。
- (d) 利用 LU 分解，解 $A\vec{x} = \vec{b}$ 。步驟：(i) 先利用 $L\vec{y} = \vec{b}$ 求出 \vec{y} 。(ii) 再利用 $U\vec{x} = \vec{y}$ 求出 \vec{x} 。(提示： $\vec{y} = L^{-1}\vec{b}$)。

5.23 大一線性代數上學期期末考有一証明題：

Let A be an $m \times n$ matrix. Prove that $\text{rank}(A) = \text{rank}(A^T A)$ 。

今天小明很無聊，想用 R 程式來驗証一下，他的做法如下：

- 先寫一 R 函式，命名為 `Check_Rank`，輸入為 `m, n`，輸出為一資料框 (`data.frame`)，欄位為英文命名，依序為：`no`(次數), `m`, `n`, `rankA`, `rankATA`。
- 在 `Check_Rank` 函式內，隨機產生任一 $m \times n$ 矩陣，隨機數來自標準常態分佈：`x <- rnorm(n*m)`。
- 查詢在 R 如何計算一矩陣的 Rank。
- 以 `seq` 指令產生 $m = 3, 5, 7, 9, 3, 5, 7, 9, 3, 5, 7, 9$ 及 $n = 4, 4, 4, 4, 8, 8, 8, 12, 12, 12, 12$ ，共 12 次組合當測試。

現在請你幫他實現這個做法。

6 機率與統計

6.1 假設一硬幣出現正面的機率為 p 。現擲此硬幣 n 次，可給出其出現正面的次數 x 的機率為 $p_x = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$, $x = 0, 1, \dots, n$ 。現假設 $n = 15$, 當 $p = 0.2$ 和 $p = 0.8$ 時，計算出 p_x 之值，如下表。

x	px0.2	px0.8
1	0.0352	0.0000
2	0.1319	0.0000
3	0.2309	0.0000
4	0.2501	0.0000
5	0.1876	0.0000
6	0.1032	0.0001
7	0.0430	0.0007
8	0.0138	0.0035
9	0.0035	0.0138
10	0.0007	0.0430
11	0.0001	0.1032
12	0.0000	0.1876
13	0.0000	0.2501
14	0.0000	0.2309
15	0.0000	0.1319
16	0.0000	0.0352

6.2 一袋中有 6 顆白球 4 顆紅球，隨機從中抽取 3 球 (取出不放回)，若 $P(A)$ 代表抽中 2 顆白球及 1 顆紅球的機率，試求 $P(A)$ 。

sol:

$$P(A) = \frac{C_2^6 C_1^4}{C_3^{10}} = \frac{1}{2}.$$

小明想要以程式方式模擬抽球來計算此機率。

- (a) 若設定 `set.seed(123456)`，列出「一袋中有 6 顆白球 4 顆紅球，隨機從中抽取 3 球」實驗一次的結果，並計數印出白球及紅球各出現之個數。
- (b) 同上小題，重覆上述實驗 10 次，計數並印出白球及紅球各出現之個數，如下。

```
> DrawResult
```

白球 紅球

1	1	2
2	2	1
3	2	1
4	2	1
5	0	3
6	1	2
7	2	1
8	2	1
9	1	2
10	1	2

(c) 同上小題，重覆上述實驗 100 次，計算抽中 2 顆白球及 1 顆紅球的機率。

6.3 一袋中有 5 顆紅球及 3 顆白球，小明由袋中隨機抽球，每次取一球，共取 4 次，令 A 為抽出 2 次白球的事件，計算此事件分別在放回 (replacement)、不放回 (without replacement) 兩種情況下之機率 $P(A_r), P(A_w)$ 。

Solution:

$$\text{放回} : P(A_r) = C_2^4 \left(\frac{3}{8}\right)^2 \left(\frac{5}{8}\right)^2$$

$$\text{不放回} : P(A_w) = \frac{C_2^5 C_2^3}{C_4^8}$$

- (a) 請直接計算上述之機率 (分別命名為 Prob.Ar, Prob.Aw) 並印出。(註: C 為組合數，R 指令為 choose)
- (b) 小明今天想要以寫 R 程式的方式來模擬此隨機實驗，計算抽球的機率。若設定 set.seed(123456)，列出「一袋中有 5 顆紅球及 3 顆白球，小明由袋中隨機抽球，分別在放回 (replacement)、不放回 (without replacement) 兩種情況下，每次取一球，共取 4 次」實驗一次的結果，並計數印出自球出現之個數。(不需寫成 R 函式)
- (c) 同上小題，寫一 R 函式 (命名 Draw_Result)，輸入為「是否印出每次實驗結果 (is.print)」，預設值為 TRUE，輸出為在兩種情況下，白球各自出現之個數。印出重覆此實驗 10 次的結果。(提示: as.data.frame, replicate)

```
> Draw.10.result
白球_放回 白球_不放回
 1      4      1
 2      2      2
 3      1      2
 4      2      0
 5      1      2
 6      3      2
 7      3      1
 8      1      2
 9      2      2
10      0      1
```

- (d) 同上小題，重覆上述實驗 1000 次，不印出每次結果，分別在放回、不放回兩種情況下，計算抽中 2 顆白球機率。(提示: `as.data.frame`, `replicate`, `sum`, `==`)

6.4 袋中有 r 顆紅球、 w 顆白球，從中每次任取一球，取後放回，共取 k 次，則 k 次取球中恰取得 p 次紅球的機率為何？試寫一 R 函式 (命名為 `draw_ball_prob`)，計算此抽球機率。執行程式時，以 $r = 3, w = 4, k = 6, p = 2$ 為例。

6.5 丟一顆公平的骰子 (正六面體)100 次，計算每個數字出現的次數。

6.6 擲兩個公平的骰子 100 次，列出各點數和之出現次數。(提示如下)

```
set.seed(12345)
n <- 100
dice.1 <- sample(1:6, n, replace=T)
...
```

6.7 擲一公平骰子，結果為 (1,2,3,4,5,6) 其中一個。令隨機變數 X_i 為擲 n 顆骰子中，數字 i 出現之次數，其符合多項式分佈。(以下各題假設從多項式分佈中隨機抽樣)

- (a) 假設丟 $n = 10$ 顆骰子 1 次，計算其「平均點數」。
- (b) 假設丟 $n = 10$ 顆骰子 2 次，計算其「平均點數」之平均。
- (c) 寫一 R 副函式 (function)，計算 m 次實驗中，「平均點數」之樣本平均值。(注意：其輸入參數為 m, n)

6.8 有兩顆骰子，一顆是公正的 (即出現 1 點 ~6 點的機率是一樣的)，另一顆不是公正的 (其奇數點出現的機率是偶數點的兩倍)。小明同時丟這兩顆骰子 100 次。(Hint: `set.seed(12345)`)

- (a) 請畫出兩個骰子出現點數之散佈圖，其中 x 軸為公正骰子出現點數， y 軸為不公正骰子出現點數，散佈圖上的點，填上色階，代表出現之次數，要有圖例說明 (legend)。
- (b) 請列出兩個骰子點數和之分佈。
- (c) 請畫出兩個骰子點數和之長條圖。

6.9 請用 R 模擬算機率：擲二個公平的六面體骰子，出現點數和為 8 的機率是多少？
(理論值為 $\frac{5}{36} = 0.1388889$)

6.10 大樂透為 1~42 個號碼一次抽出 6 個號碼為一組 (取出不放回)。請模擬抽獎 100 次 (組)，並計算每個數字 (1~42) 出現的次數。

6.11 有一副牌共編號 1~100。均勻洗牌後，便依序逐一翻牌，並從 1 數到 100，若數的號碼與牌上的編號相同時，便稱為一個 hit。

- (a) 請寫一 R 函式 (命名為 `hit_no`)，利用 `sample` 和邏輯判斷，寫出 hit 數。(提示：輸入為編號個數 100，輸出為 hit 個數。)
- (b) 可知 hit 數為一隨機變數，模擬玩 1000 次，計算此 1000 次的平均數和標準差。(提示：`replicate`)

6.12 自一副標準的 52 張撲克牌中，隨機取 5 張牌，現想估計此 5 張牌中至少有一張牌為 8 點之機率。請依下列步驟和提示完成。(四種花色：黑桃 (spade)、紅心 (heart)、方塊 (diamond)、梅花 (club))

- (a) 先造一 `poker` 的 `data.frame` 如下：

```
> poker
  suit.cht number
 1    黑桃     1
 2    黑桃     2
 3    黑桃     3
 4    黑桃     4
 5    黑桃     5
 6    黑桃     6
 7    黑桃     7
 8    黑桃     8
 9    黑桃     9
10   黑桃    10
11   黑桃    11
12   黑桃    12
13   黑桃    13
14   紅心     1
15   紅心     2
16   紅心     3
17   紅心     4
18   紅心     5
...
48   梅花     9
49   梅花    10
50   梅花    11
51   梅花    12
52   梅花    13
```

- (b) 利用此 `poker` 隨機抽取 5 張牌。再印出 5 張牌中有 8 點的張數。
- (c) 重覆隨機取 5 張牌的過程 1000 次，每次計算 5 張牌中有 8 點的張數，印出「張數的次數表」。(提示: `table`)

6.13 一副撲克牌共有 52 張牌，在撲克牌遊戲中常會以五張牌的組合，比較大小來決定勝負。若隨機抽取 5 張牌，取後放回，可分成下列 10 不同的撲克牌型：

牌型	英文	組合	機率
同花順	straight flush	40	0.0015391%
四條、鐵支	four of a kind	624	0.0240096%
葫蘆、滿堂紅	full house	3,744	0.1440576%
同花	flush	5,108	0.1965402%
順子	straight	10,200	0.3924647%
三條	three of a kind	54,912	2.1128451%
兩對	two pair	123,552	4.7539016%
一對	one pair	1,098,240	42.2569028%

- (a) 印出上表。
- (b) 在 `set.seed(1234567)` 之下，重覆「隨機抽取 5 張牌」100000 次，列出上述這些牌型共出現幾次？
- (c) 利用 `repeat`，計算出要取到第幾次才會出現兩對（兩組不同點數的牌各兩張）。
- (d) 利用 `while`，計算出要取到第幾次才會出現兩對。
- (e) 在出現葫蘆（三張點數相同的牌和另外兩張同點數的牌）之前，兩對的情況共出現幾次？

6.14 根據人口普資料，美國成年人口的分類比例如下：

類別	年齡	比例
1	18–24	0.18
2	25–34	0.23
3	35–44	0.16
4	45–64	0.27
5	65–	0.16

隨機選出 5 位成人，試問其中有 1 位介於 18–24 歲，2 位介於 25–34，2 位介於 45–64 的機率為何？

<解> 因選出的人數 5 位，遠小於全國總成人數，故可視為放回、獨立、相同之試驗，重覆 5 次。此為多項式分佈如下：

$$(Y_1, Y_2, Y_3, Y_4, Y_5) \sim \text{Multinomial}(n, p_1, p_2, p_3, p_4, p_5),$$

其中 $n = 5, p_1 = 0.18, p_2 = 0.23, p_3 = 0.16, p_4 = 0.27, p_5 = 0.16$. 所求為 $P(1, 2, 0, 2, 0) = 0.0208$.

請利用 R 程式（指令）直接計算上述之機率。

6.15 公式

$$z = \frac{x - \bar{x}}{s} \quad (\text{其中 } \bar{x} \text{ 為平均數， } s \text{ 為標準差})$$

一般稱做 z -轉換 (z -transformation) 或 z -分數 (z -score)。

- (a) 請寫一函式，輸入為一數列，輸出為 z 分數。

- (b) 若有 5 個成績 `x <- sample(0:100, 5)`，請利用上述之函式轉成 z 分數，並算出轉換後之平均數及變異數。
- (c) 請與 `scale` 之結果相比較。

6.16 常態分佈的機率密度函數如下：

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- (a) 請寫一 R 函式 (命名為 `my.dnorm`)，計算常態分佈的機率密度函數值，輸入為 x 值、平均數 μ (預設值為 0) 及標準差 σ (預設值為 1)，輸出為常態分佈的機率密度函數值 $f(x; \mu, \sigma)$ 。使用 `my.dnorm` 計算 $f(2.5; 3, 2)$ 之值。
- (b) R 內建計算常態分佈的機率密度函數值的指令為 `dnorm`，印出下列表格 (標準常態分佈)：

x	my.dnorm	dnom
1 -3	0.004431848	0.004431848
2 -2	0.053990967	0.053990967
3 -1	0.241970725	0.241970725
4 0	0.398942280	0.398942280
5 1	0.241970725	0.241970725
6 2	0.053990967	0.053990967
7 3	0.004431848	0.004431848

6.17 (卜瓦松累積機率分佈函數) 若隨機變數 Y 服從卜瓦松分配 ($X \sim Poisson(\lambda)$)，其累積機率分佈函數為：

$$F_X(x) = P(X \leq x) = \sum_{k=0}^x \frac{\lambda^k e^{-\lambda}}{k!}$$

試寫一 R 函數 (命名為 `poisson_cdf`，輸入為 x 、 λ ，輸出為 $F_X(x)$)，計算卜瓦松分配之累積機率分佈函數值，並利用此函數印出下表 (印出紅色框部份的表格即可)。

Tables of the Poisson Cumulative Distribution

The table below gives the probability of that a Poisson random variable X with mean = λ is less than or equal to x . That is, the table gives

$$P(X \leq x) = \sum_{r=0}^x \lambda^r \frac{e^{-\lambda}}{r!}$$

$\lambda =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4	1.6	1.8	
$x =$	0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3012	0.2466	0.2019	0.1653
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6626	0.5918	0.5249	0.4628	
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.8795	0.8335	0.7834	0.7306	
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9903	0.9865	0.9810	0.9662	0.9463	0.9212	0.8913	
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9980	0.9977	0.9963	0.9923	0.9857	0.9763	0.9636	
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9994	0.9985	0.9968	0.9940	0.9896	
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9974	
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9994	
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
$\lambda =$	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.5	5.0	5.5	
$x =$	0	0.1353	0.1108	0.0907	0.0743	0.0608	0.0498	0.0408	0.0334	0.0273	0.0224	0.0183	0.0111	0.0067	0.0041
1	0.4060	0.3546	0.3084	0.2674	0.2311	0.1991	0.1712	0.1468	0.1257	0.1074	0.0916	0.0611	0.0404	0.0266	
2	0.6767	0.6227	0.5697	0.5184	0.4695	0.4232	0.3799	0.3397	0.3027	0.2684	0.2381	0.1736	0.1247	0.0884	
3	0.8571	0.8194	0.7787	0.7360	0.6919	0.6472	0.6025	0.5584	0.5152	0.4735	0.4335	0.3423	0.2650	0.2017	
4	0.9473	0.9275	0.9041	0.8774	0.8477	0.8153	0.7806	0.7442	0.7064	0.6678	0.6288	0.5321	0.4405	0.3575	
5	0.9834	0.9751	0.9643	0.9510	0.9349	0.9161	0.8946	0.8705	0.8441	0.8156	0.7851	0.7029	0.6160	0.5289	
6	0.9955	0.9925	0.9884	0.9828	0.9756	0.9665	0.9554	0.9421	0.9267	0.9091	0.8893	0.8311	0.7622	0.6860	
7	0.9989	0.9980	0.9967	0.9947	0.9919	0.9881	0.9832	0.9769	0.9692	0.9599	0.9489	0.9134	0.8666	0.8095	
8	0.9998	0.9995	0.9991	0.9985	0.9976	0.9962	0.9943	0.9917	0.9883	0.9840	0.9786	0.9597	0.9319	0.8944	
9	1.0000	0.9999	0.9998	0.9996	0.9993	0.9989	0.9982	0.9973	0.9960	0.9942	0.9919	0.9829	0.9682	0.9462	
10	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992	0.9987	0.9981	0.9972	0.9933	0.9863	0.9747	
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9996	0.9994	0.9991	0.9976	0.9945	0.9890	
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9992	0.9980	0.9955	
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9983	
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

6.18 (變數變換) 已知隨機變數 X_1, X_2 的聯合機率密度函數為

$$f_{X_1, X_2}(x_1, x_2) = e^{-x_1-x_2}, \quad x_1 > 0, x_2 > 0.$$

試以 R 程式舉例說明隨機變數 $Y = \frac{X_1}{X_1 + X_2}$ 是在區間 $[0, 1]$ 之連續均勻分佈。

6.19 兩變數之 n 個觀察值記為 $\{(x_i, y_i); i = 1, \dots, n\}$ ，其樣本相關係數之公式如下：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- (a) 試寫一 R 函式，計算兩變數之相關係數。
- (b) 若兩變數之資料如下：`n <- 20; x <- rnorm(n); y <- runif(n)`，試用上題之函式計算相關係數，並與 `cor` 之結果相比較。

6.20 斯皮爾曼等級相關係數 (Spearman's rank correlation coefficient) 之公式如下：

$$\rho = \frac{\sum_{i=1}^n (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R_{x_i} - \bar{R}_x)^2} \sqrt{\sum_{i=1}^n (R_{y_i} - \bar{R}_y)^2}},$$

其中 R_{x_i} 是 x_i 的 Rank(等級), R_{y_i} 是 y_i 的 Rank(等級)， (\bar{R}_x, \bar{R}_y) 是資料 (R_{x_i}, R_{y_i}) , $i = 1, \dots, n$ 之平均數。試寫一 R 函式，輸入為兩變數資料，輸出為兩變數之斯皮爾曼等級相關係數。以 `x <- iris[, 1]; y <- iris[, 3]` 為例。並與 `cor` 之結果相比較。(提示: `rank`)

6.21 一單變量資料 $\{x_i\}_{i=1}^n$ 之樣本偏態 (skewness) 係數公式如下:

$$b_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}},$$

其中 n 為資料觀察個數， \bar{x} 為此資料之平均數。試寫一 R 函式 (命名為 skewness)，計算以下資料 x 之樣本偏態係數:

```
set.seed(12345)
x <- rnorm(100)
```

6.22 峰度係數 k_c (coefficient of kurtosis) 為一測量峰度高低的量數，可以反映資料的分佈形狀。峰度係數一般是與常態分配作比較而言，該資料分配是否比較高聳或是扁平的形狀。其判別如下:

- 若 $k_c > 0$, 表示資料分布呈高狹峰 (lepto kurtosis)。
- 若 $k_c = 0$, 表示資料分布呈常態峰 (normal kurtosis)。
- 若 $k_c < 0$, 表示資料分布呈低潤峰 (platy kurtosis)。

常用的樣本峰度係數的計算式有以下三項:

- The typical definition used in many older textbooks: $g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$
- Used in SAS and SPSS: $G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$
- Used in MINITAB and BMDP: $b_2 = (g_2 + 3)\left(1 - \frac{1}{n}\right)^2 - 3$

其中 n 為樣本大小， x_i 為第 i 個測量值， \bar{x} 為平均數。

- (a) 請寫一 R 函式 (my.kurtosis)，輸入為一組學生成績 (score)，輸出為此資料的三項樣本峰度係數。

```
> set.seed(123456)
> score <- rt(150, 4)
> my.kurtosis(score)
$kc
      g2        G2        b2
1.980622 2.089356 1.914436
```

- (b) 讀入資料 score-data.txt 命名為 my.score 物件，使得欄位名稱為科目名，列名稱為學號。利用 apply 及 my.kurtosis 求每一科目的三項樣本峰度係數。(若有 NA，請以 0 分計)

```

> my.score <- ....
> apply(....)
$線代
$線代$kc
      g2          G2          b2
-0.6848024 -0.6282452 -0.7764842
...
...

```

6.23 截尾平均數 (trimmed mean) 是將一數值資料排序後，將頭尾各拿掉一定百分比 p 的觀察值，然後用剩下的觀察值計算平均數。

(a) 試寫一 R 函式 (命名為 `trim_mean`)，計算截尾平均數，輸入為一數字向量 x 及截尾比例 p ，輸出為截尾平均數。

(b) 若有兩資料 `data1` 和 `data2` 如下：

```

data2 <- data1 <- rnorm(100)
id <- sample(100, 10)
data2[id] <- data1[id] + 2*qchisq(0.975, 10)

```

呈上題，計算兩資料之截尾百分比 p 為 0%、1%、3%、5% 及 10% 之截尾平均數。

(c) 請與 `mean(data1, trim = p)`, `mean(data2, trim = p)`, 其中 $p=0, 0.01, 0.03, 0.05, 0.1$ 的結果相比較。

6.24 統計學中，”Cramér's V” 是用來量測兩個類別變數 (X, Y) 的相關程度，其值介於 0 到 +1 之間，公式如下：

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}},$$

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n})^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}},$$

其中

- n 筆觀察資料記為 $\{X, Y\}_{t=1}^n = \{(x, y) : x \in \{A_1, A_2, \dots, A_r\}, y \in \{B_1, B_2, \dots, B_k\}\}$ 。
- n : 觀察資料總個數。
- n_{ij} : 變數 X 觀察值為 i 類別且變數 Y 觀察值為 j 類別之資料個數。(提示：`table(X, Y)`)
- $n_{i\cdot}$: 變數 X 觀察值為 i 類別之資料個數。(提示：`rowSums`)
- $n_{\cdot j}$: 變數 Y 觀察值為 j 類別之資料個數。(提示：`colSums`)

- 類別變數 X 具有 r 個類別，類別變數 Y 具有 k 個類別。(提示: `length(unique(X))`)

若觀察之資料如下，試寫一 R 函式 (命名為 `myCramerV`)，輸入為兩類別變數 (X, Y)，計算其 V 值，並與 `cramer.v` {`questionr`} 相比較。(提示: (1) 使用 `for` 雙迴圈計算 χ^2 。(2) `cramer.v(table(X, Y))`)

```
set.seed(12345)
X <- sample(paste0("A", 1:4), 50, replace=T)
Y <- sample(paste0("B", 1:6), 50, replace=T)
```

6.25 計算名目變數 (nominal variable) 的變異分散程度，其中 Index of Qualitative Variation (IQV) 是一個指標 (其數值是介於 0 與 1 中間)。公式如下：

$$IQV = \frac{k(n^2 - \sum f^2)}{n^2(k - 1)},$$

其中 k 是類別數或組數， n 是樣本數， $\sum f^2$ 是將各類別次數之平方加起來之總和。假設有一名目變數資料 (`nv`) 如下，試寫一 R 函式，計算 IQV。(提示: `table`)

```
set.seed(12345)
no <- sample(20:100, 1)
nv <- LETTERS[sample(1:26, 5)][sample(1:5, no, replace=T)]
```

6.26 將所觀察到兩變數的資料記做 $\{x_i, y_i\}_{i=1}^n$ ，並進行簡單線性迴歸分析。簡單線性迴歸中 $(y = \beta_0 + \beta_1 x + \epsilon)$ 之斜率項 (β_1) 及截距項 (β_0) 的估計量如下：

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \bar{x} \text{ 和 } \bar{y} \text{ 為 } x_i \text{'s 和 } y_i \text{'s 的平均。}$$

- 試寫一 R 函式，輸入為兩變數的資料，輸出為斜率項及截距項的估計量。
- 兩變數之資料如下: `x <- iris[,1]; y <- iris[,2]`，試用上題之函式計算斜率項及截距項的估計量，並與 `lm(y~x)` 之結果相比較。

6.27 將所觀察到兩變數 (x, y) 的 n 筆紀錄記做 $\{x_i, y_i\}_{i=1}^n$ ，並將此資料進行簡單線性迴歸分析，其中 y 為反應變數， x 為解釋變數。簡單線性迴歸中 $(y = \beta_0 + \beta_1 x + \epsilon)$ 之斜率項 (β_1) 及截距項 (β_0) 的估計量如下：

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, \bar{x} 和 \bar{y} 為 x_i 's 和 y_i 's 的平均。

假設兩變數之資料如下: `x <- iris[,1]; y <- iris[,2]` .

- (a) 利用上述公式，試計算斜率項及截距項的估計量。
- (b) 以矩陣型態表示斜率項 (β_1) 及截距項 (β_0) 的估計量如下:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^T X)^{-1} X^T Y, \text{ 其中 } X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

利用矩陣表示法求 $\hat{\beta}_1$ 及 $\hat{\beta}_0$ 。

- (c) 計算結果請與 `lm(y~x)` 之結果相比較。

6.28 獨立雙樣本 t 檢定 (Two-sample t-test) 是用來檢定兩母體之平均數是否相同，其虛無假設為:

$$H_0: \mu_x = \mu_y.$$

假設兩母體之變異數不相等之下，從中所抽取的兩組樣本 $\{x_1, x_2, \dots, x_n\}$, $\{y_1, y_2, \dots, y_m\}$ 其 t 檢定統計量公式為

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}},$$

其中 \bar{x} 為樣本 x_i 's 的平均數， s_x^2 為樣本 x_i 's 的變異數。

- (a) 若某生想檢定兩樣本 `x <- iris[,2]; y <- iris[,3]` 之母體平均數是否相等，請你用 R 指令 `t.test` 幫他完成檢定。
- (b) 在虛無假設之下，t 檢定統計量服從 t 分佈，具有自由度

$$df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)}.$$

此檢定之 p 值 (p-value) 為 $P(T > |t|)$ 。兩母體平均差 $\mu_x - \mu_y$ 之 $(1 - \alpha)\%$ 信賴區間為

$$CI = (\bar{x} - \bar{y}) \pm t_{1-\alpha/2, df} \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}.$$

請寫一 R 函數 (命名為 `my.t`)，利用上小題之資料，計算並印出相關資訊如下。

```
> my.t(x, y)
My Two Sample t-test

Sample means of x and y:  3.057333      3.758
t = -4.719421 , df = 167.0999 , p-value = 4.975454e-06
95 percent confidence interval: -0.9937746 -0.4075587
```

6.29 寫一 R 函數，檢定「兩常態母體變異數相等」(F-test, 以雙尾檢定為例)。

$$x_1, \dots, x_m \sim N(\mu_1, \sigma_1^2), y_1, \dots, y_n \sim N(\mu_2, \sigma_2^2), H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1.$$

(a) 輸入：資料 x, y 。顯著水準 (α)。

輸出：(1) 標題。(2) 檢定統計值： $(f = \frac{S_1^2}{S_2^2})$ 。(3) 自由度： $(m - 1, n - 1)$ 。

(4) 臨界值： $(F_{1-\alpha/2, m-1, n-1}, F_{\alpha/2, m-1, n-1})$ 。(5) $(1-\alpha)\%$ 信賴區間： $(P(\frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha/2}}) = 1 - \alpha)$ 。(6) p-value： $2 \times P(F \geq f)$ 。(7) 決策。

(b) 用以下例子做測試，並和 `var.test` 做比較。某施工現場，使用兩種類堆土機，已知 B 機種的能力較 A 機種為優秀。A、B 兩種堆土機進行 6 天的挖掘工作，比較其能力 ($m^3/\text{日}$) 如下：

A: 68.8, 65.7, 67.6, 67.8, 66.2, 66.8

B: 69.0, 68.2, 69.4, 67.1, 68.8, 68.2

試問如 B 機種之變異程度和 A 機種相當嗎？($\alpha = 0.05$)。

6.30 卡方獨立性檢定適用於分析兩組類別變數 (X, Y) 是否為相互獨立。例如：薪資的高低和學歷的程度是否獨立？婚姻狀況與宗教信仰是否有關？通常資料會表示成一 $r \times c$ 的列聯表 (Contingency Table)(格式如下)，其中 $\{A_1, A_2, \dots, A_r\}$, $\{B_1, B_2, \dots, B_c\}$ 分別為兩組類別變數 (X, Y) 之類別，其中 O_{ij} 為 $(X = A_i, Y = B_j)$ 觀察次數， n 為樣本數：

(X, Y)	B_1	B_2	\cdots	B_j	\cdots	B_c	列總和
A_1	O_{11}	O_{12}	\cdots	O_{1j}	\cdots	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	\cdots	O_{2j}	\cdots	O_{2c}	$O_{2\cdot}$
\vdots	\vdots			\vdots		\vdots	\vdots
A_i	O_{i1}	O_{i2}	\cdots	O_{ij}	\cdots	O_{ic}	$O_{i\cdot}$
\vdots	\vdots			\vdots		\vdots	\vdots
A_r	O_{r1}	O_{r2}	\cdots	O_{rj}	\cdots	O_{rc}	$O_{r\cdot}$
行總和	$O_{\cdot 1}$	$O_{\cdot 2}$	\cdots	$O_{\cdot j}$	\cdots	$O_{\cdot c}$	n

卡方獨立性檢定的統計值公式為

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

其中 E_{ij} 為期望次數，計算方式是以第 i 列總和與第 j 行總和之乘積除以總樣本數 n 而得：

$$E_{ij} = \frac{O_{i\cdot} \times O_{\cdot j}}{n}.$$

現有一資料如下，想了解性別與政黨傾向是否獨立。

Gender	Party Identification		
	Democrat	Independent	Republican
Females	762	327	468
Males	484	239	477

- (a) 將上述資料表格，輸入 R 成為一矩陣 (matrix) 類別的物件後，將之轉成表格 table 類別物件 (命名為 `GenderParty.observed.ct`)，並加上維度名稱後印出如下。(提示: `as.table`, `dimnames`)

```
> GenderParty.observed.ct
      party
gender Democrat Independent Republican
  F       762          327        468
  M       484          239        477
```

- (b) 計算並印出列總和、行列總和及樣本總個數。
(c) 計算並印出期望次數表格 (命名為 `GenderParty.expected.ct`) (注意，此物件之 R 類別同 `GenderParty.observed.ct`)。
(d) 計算並印出此資料的卡方獨立性檢定統計值 χ^2 。

6.31 若隨機變數 X 服從二項式分配 ($X \sim Binomial(n, p)$)，其機率質量函數為：

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x},$$

試寫一 R 函數 (命名為 `binomial.pmf`)，計算二項式分配之機率質量函數值。

6.32 若隨機變數 Y 服從卜瓦松分配 ($Y \sim Poisson(\lambda)$)，其機率質量函數為：

$$f_Y(y) = P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}.$$

試寫一 R 函數 (命名為 `poisson.pmf`)，計算卜瓦松分配之機率質量函數值。

6.33 If the random variable X follows the binomial distribution with parameters $n \in N$ and $p \in [0, 1]$, we write $X \sim B(n, p)$. The probability of getting exactly k successes in n trials is given by the probability mass function (pmf):

$$f(k; n, p) = Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}.$$

The expected value (mean) and variance of X are np and $np(1 - p)$, respectively. The formulas are:

$$\mu = \sum_{k=0}^n kf(k; n, p), \quad \text{and} \quad \sigma^2 = \sum_{k=0}^n (k - \mu)^2 f(k; n, p).$$

試寫一 R 函數，以數值計算二項式分佈隨機變數 ($X \sim B(20, 0.3)$) 之平均數及變異數。

```

binomial <- function(k, n, p){
  ...
}

compute.mu.sigma <- function(pmf, parameter){
  ...
  cat("mu: ", mu, "\n")
  cat("sigma2: ", sigma2, "\n")
}

n <- 20
p <- 0.3
k <- ...
my.knp <- list(k, n, p)
compute.mu.sigma(pmf=binomial, parameter=my.knp)
mu: ...
sigma2: ...

```

6.34 在日常生活中，我們可以找到很多具有二元反應的例子。例如：銅板有正面、反面。對某議題你可能會支持或不支持。在還未表態之前，我們不會知道會出現什麼結果，但我們又很想對這個情況作一個描述，因此發展出機率來解釋這些現象。首先，令這隨機現象出現的結果以 X 表示，其二元反應結果值為 0 或 1，機率表示為 $P(X = 1) = 1 - P(X = 0) = p$, $0 \leq p \leq 1$ (這裏的大寫表示是隨機的，就是還沒看到結果，有可能為 0 或 1)。又當你擲銅板 n 次、或是問了 n 個人是否支持，其結果則以 X_1, X_2, \dots, X_n 來表示，令 $Y = \sum_{i=1}^n X_i$ 為 n 次結果中，出現 1 的次數和，可能值為 $0, 1, \dots, n$ 。 Y 是一個隨機變數，其機率質量函數表示為

$$P(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \quad y = 0, 1, \dots, n.$$

假設 $n = 15, p = 0.5$:

- (a) 這裏我們會說 X 服從 Bernoulli 分配，以 $X \sim Ber(p)$ 表示之； Y 服從 Binomial 分配，以 $Y \sim B(n, p)$ 表示之(則 $B(1|p) = Ber(p)$)。畫出 $B(n, p)$ 之機率圖形(橫軸為 y 值，縱軸為機率值 $P(Y = y)$)。(提示：使用指令 dbinom)

- (b) Y 的累積機率函數為 $P(Y \leq y) = \sum_{i=1}^{\lfloor y \rfloor} P(Y = i)$ ，其中 $\lfloor y \rfloor$ 是小於或等於 y 的最大整數。畫出 Y 之累積機率圖形(橫軸為 y 值，縱軸為累積機率值

$P(Y \leq y))$ 。(提示: 使用指令 pbinom)

(c) (承上兩小題) 列出 $B(n = 15, p = 0.5)$ 的機率質量函數、累積機率函數如下:

x	pmf	cdf
1	0.00003	0.00003
2	0.00046	0.00049
3	0.00320	0.00369
4	0.01389	0.01758
5	0.04166	0.05923
6	0.09164	0.15088
7	0.15274	0.30362
8	0.19638	0.50000
9	0.19638	0.69638
10	0.15274	0.84912
11	0.09164	0.94077
12	0.04166	0.98242
13	0.01389	0.99631
14	0.00320	0.99951
15	0.00046	0.99997
16	0.00003	1.00000

(d) 利用上述 $B(n, p)$ 的機率質量函數及累積機率函數公式，各寫出一 R 函式，命名為 my.dbinom 及 my.pbinom，並與上小題之結果相比較。

6.35 以下三個分佈是機率論及統計學中常見的離散機率分佈，其機率質量函數 (probability mass function, pmf)、期望值和變異數如下表所列。

分佈名稱	簡記	pmf ($P(X = k)$)	Support (\mathcal{S})	期望值	變異數
二項式分佈	$B(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, 1, 2, \dots, n$	np	$np(1-p)$
卜瓦松分佈	$Poi(\lambda)$	$\frac{e^{-\lambda} \lambda^k}{k!}$	$k = 0, 1, 2, \dots$	λ	λ
幾何分布	$G(p)$	$(1-p)^k p$	$k = 0, 1, 2, \dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

若隨機變數 X 服從某一分布，其期望值及變異數的計算公式如下：

$$\mu = \sum_{k \in \mathcal{S}} k P(X = k), \quad \text{and} \quad \sigma^2 = \sum_{k \in \mathcal{S}} (k - \mu)^2 P(X = k).$$

試寫一 R 函數，以數值計算

- (a) 二項式分佈隨機變數 $X \sim B(10, 0.6)$;
- (b) 卜瓦松分佈隨機變數 $X \sim Poi(4)$;
- (c) 幾何分佈隨機變數 $X \sim G(0.4)$,

之期望值及變異數。(提示: $k = 0, 1, 2, \dots, 100$)。程式提示:

```
binomial <- function(k, n, p){
  ...
}

poisson <- function(k, lambda){
  ...
}

geometric <- function(k, p){
  ...
}

compute.mu.sigma <- function(pmf, parameter){
  ...
  distribution <- deparse(substitute(pmf))
  ...
  cat("distribution: ", distribution, "\n")
  cat("mu: ", mu, "\t sigma2: ", sigma2, "\n")
}
```

```
#(a)
k <- ...
my.par <- list(k=..., n=10, p=0.6)
compute.mu.sigma(pmf=binomial, parameter=my.par)
distribution: binomial
mu: ... sigma2: ...
```

```
#(b)
my.par <- list(k=..., lambda=...)
compute.mu.sigma(pmf=binomial, parameter=my.par)
distribution: poisson
mu: ... sigma2: ...
```

```
#(c)
my.par <- list(k=..., p=...)
compute.mu.sigma(pmf=binomial, parameter=my.par)
distribution: geometric
mu: ... sigma2: ...
```

6.36 若 (X, Y) 為服從二元常態分佈 (The Bivariate Normal Distribution) 之隨機變數，其聯合機率密度函數 (joint probability density function) 如下：

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right] \right).$$

其中參數 (μ_X, μ_Y) 為 (X, Y) 之平均數， (σ_X^2, σ_Y^2) 為 (X, Y) 之變異數， ρ 為 X 和 Y 之相關係數。我們以下列符號表示：

$$(X, Y) \sim BVN(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

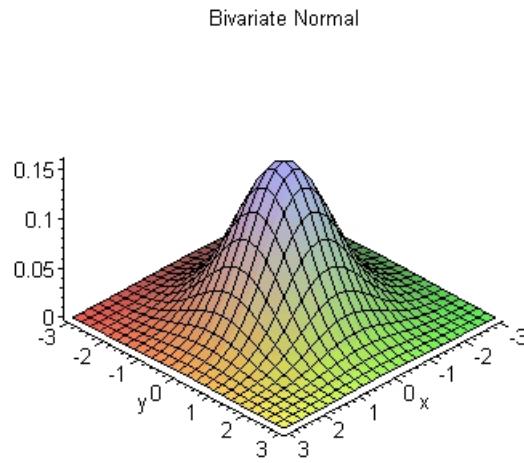
- (a) 試寫一 R 函數 (命名為 `my.dbvn`)，計算二元常態分佈聯合機率密度函數值 (R 函數之輸入為 (x, y) 及上述 5 個參數)。
- (b) 若 $(X, Y) \sim BVN(1, 0, 2, 0.5, 0.6)$ ，計算二元常態分佈機率密度函數值，如下表：

	x	y	fxy
1	-1	-2	0.003423
2	-1	-1	0.056998
3	-1	0	0.041701
	...		
23	3	0	0.041701
24	3	1	0.056998
25	3	2	0.003423

(提示：)

```
x <- seq(-1, 3, 1)
y <- seq(-2, 2, 1)
xy <- data.frame(x=rep(x, each=length(y)), y=rep(y, length(x)))
fxy <- ...
...
...
```

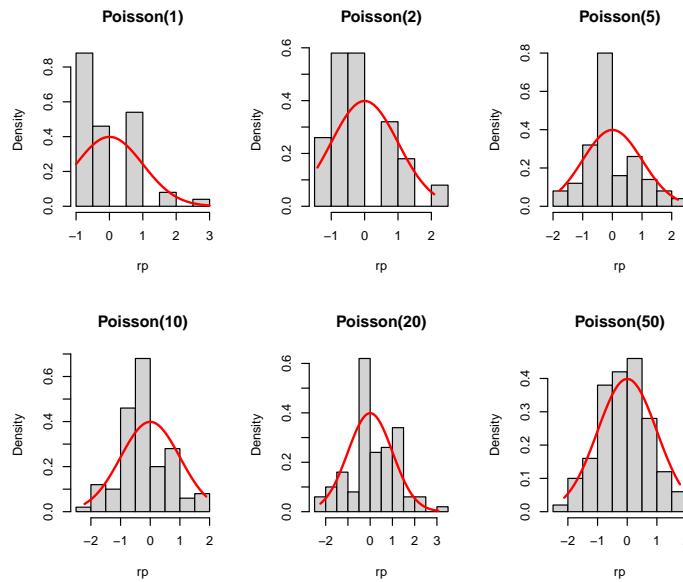
- (c) (承上小題) 請將結果與 `dmvnorm` {mvtnorm} (R 版本 >3.5) 或 `dmvn` {LaplaceDemon} 相比較。(提示：意即利用 `dmvnorm` {mvtnorm} 或 `dmvn` {LaplaceDemon} 印出上小題表格。)
- (d) 利用 `my.dbvn`，以 `persp` {graphics} 繪出標準二元常態分佈聯合機率密度函數圖。(曲面顏色可用 `tim.colors` {fields}) (範例如下：<http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html>) (提示：`outer`)



6.37 以常態分佈逼近布瓦松 (Poisson) 分佈 (Normal Approximation to Poisson Distribution):

$$\text{If } X \sim \text{Poisson}(\lambda) \text{ then } \frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{d} \text{Normal}(0, 1) \text{ for a sufficient large } \lambda.$$

使用 $\lambda = 1, 2, 5, 10, 20, 50$ 重覆下列步驟來驗証。(共 6 個圖，請畫成一頁 2×3 , set.seed(123456)): (1) 隨機產生 100 個 $\text{Poisson}(\lambda)$ 隨機數，將資料利用 $\frac{x-\lambda}{\sqrt{\lambda}}$ 轉換後，畫出其直方圖 (圖標題是 $\text{Poisson}(\lambda)$ ， λ 需換成數字)。 (2) 在直方圖上加上 (紅色) 標準常態分佈曲線。

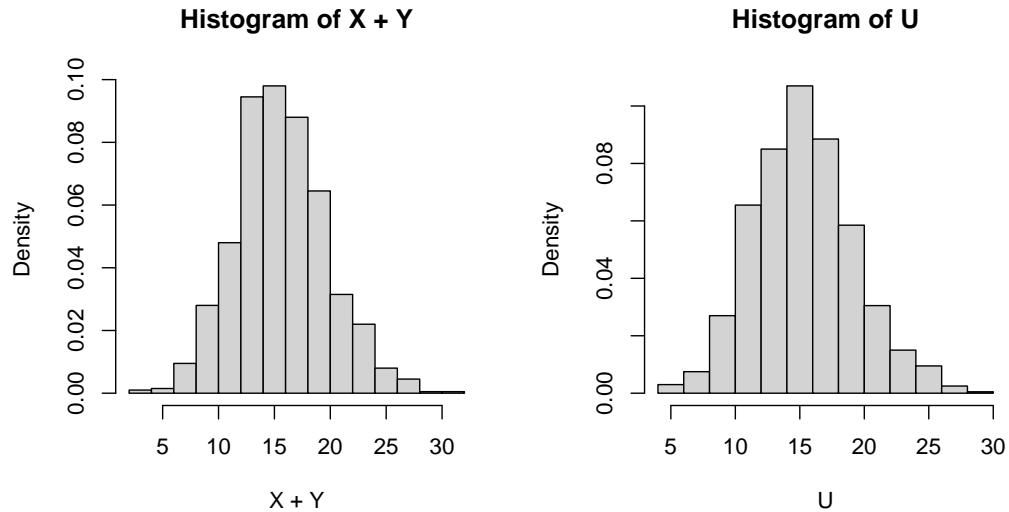


6.38 請依下列兩方式，利用 R 程式驗証下述定理。

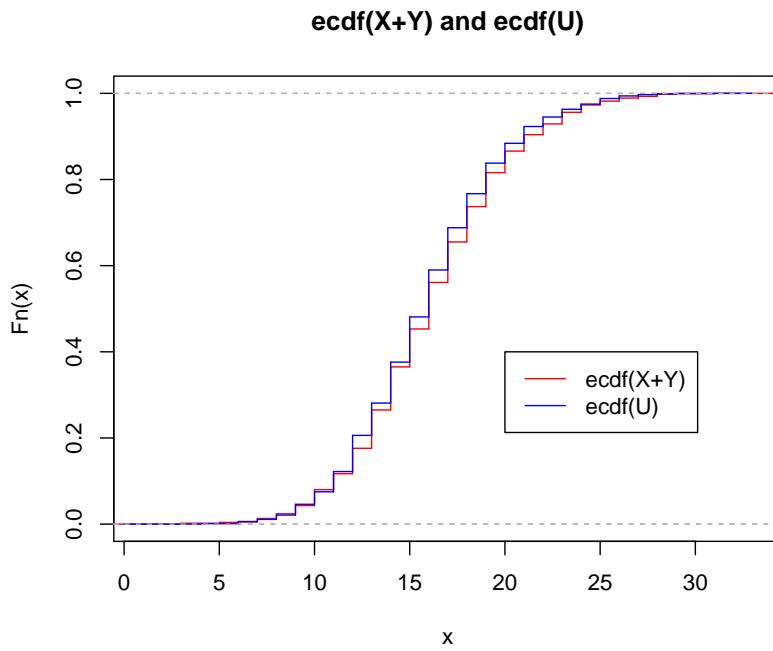
Theorem: Let $X \sim \text{Poisson}(\theta)$, $Y \sim \text{Poisson}(\lambda)$ and X and Y are independent. If $U = X + Y$, then $U \sim \text{Poisson}(\theta + \lambda)$.

- (a) 由 $\text{Poisson}(\theta = 5)$ 產生 1000 個隨機數，令此為隨機變數 X 之實現值。再由 $\text{Poisson}(\lambda = 11)$ 產生 1000 個隨機數，令此為隨機變數 Y 之實現值。

產生 $Poisson(\theta + \lambda = 5 + 11)$ 的隨機數 1000 個，令此為隨機變數 U 之實現值。畫出 $X + Y$ 之直方圖與 U 之直方圖。(提示: (1) 一頁兩張圖。(2) `set.seed(12345)`。(3) `hist(..., freq=FALSE)`)



(b) (承上小題) 將 $X + Y$ 與 U 之經驗分布函數 (Empirical Cumulative Distribution Function) 繪出如下。(提示: (1) `ecdf`。(2) `plot(..., verticals = TRUE, do.points = FALSE)`)



6.39 Poisson 極限定理 (Poisson Limit Theorem):

給定常數 λ · 且 $X_n \sim Binomial(n, \frac{\lambda}{n})$, 則

$$P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \rightarrow \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{as } n \rightarrow \infty$$

亦即 · Poisson 分配是二項式分配 $Binomial(n, \frac{\lambda}{n})$ 的一個極限分配。

若 $n = 30, \lambda = 2.4$ · 利用上兩小題所寫的 R 函式 `binomial.pmf`, `poisson.pmf`, 列出下表來驗証 Poisson 極限定理, 其中最一欄位 (diff) 為 `binomial.pmf`, `poisson.pmf` 兩者差之絕對值。

x	<code>binomial.pmf</code>	<code>poisson.pmf</code>	diff
1 0	0.0820	0.0907	0.0087517497
2 1	0.2138	0.2177	0.0038982090
3 2	0.2696	0.2613	0.0083375766
4 3	0.2188	0.2090	0.0097959196
5 4	0.1284	0.1254	0.0030235072
6 5	0.0581	0.0602	0.0021224767
7 6	0.0210	0.0241	0.0030372714
8 7	0.0063	0.0083	0.0019823213
9 8	0.0016	0.0025	0.0009083534
10 9	0.0003	0.0007	0.0003270817
11 10	0.0001	0.0002	0.0000976313

6.40 有一資料紀錄 11 位女性每日能量攝取量 (daily energy intake (Kilojoules, kJ) · 單位是仟焦耳)。

5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770

(a) 畫出資料的直方圖。

(b) 檢定女性每日能量攝取量是否為 7725kJ(假設資料來自於常態分佈)。

6.41 飲用水中若鋅 (zinc) 的濃度過高會危害人體健康。某研究人員想了解水杯底部鋅的濃度是否高於水杯表面鋅濃度高 · 於是從 10 杯水中取出水杯表面 (surface) 及水杯底部 (bottom) 的水 · 各測得其鋅的濃度 · 紀錄於 `water.txt`。

(a) 畫出散佈圖 · 橫軸為水杯底部鋅濃度 · 縱軸為水杯表面鋅濃度。並在圖上加一通過原點的 45 度直線。

(b) 請你幫他做檢定。

6.42 某醫院進行藥物測驗 · 測得實驗組及對照組之指標如下:

實驗組 86, 72, 74, 85, 76, 79, 82, 83, 83, 79, 82

對照組 81, 77, 63, 75, 69, 86, 81, 60

(a) 畫出兩組之 side-by-side 盒形圖。

(b) 請檢定兩組之指標值有無顯著差異?

6.43 R 內建資料 ChickWeight {datasets}。578 小雞在成長過程中分別餵食 4 種不同的蛋白質食物 (Diet)。經過 20 天之後，量得它們的體重 (weight)。

- (a) 畫出四組之 side-by-side 盒形圖。
- (b) 請檢定四組之小雞體重有無顯著差異? (變異數分析)

6.44 以下三個 R 套件皆提供一些函數可計算 Shrinkage estimation of covariance matrix:

- cov.shrink {corpcor}
- shrinkcovmat.identity {ShrinkCovMat}
- covEstimation {RiskPortfolios} with type = 'oneparm'

現以 R 程式產生一模擬資料 x 如下

```
library(MASS)
n <- 10
p <- 100
set.seed(123456)
sigma <- matrix(rnorm(p * p), ncol = p)
sigma <- crossprod(sigma) + diag(rep(0.2, p))
x <- mvrnorm(n, mu=rep(0, p), Sigma=sigma)
```

試以不同的 p/n 值 ($p/n = 0.1, 0.5, 2, 10$, p 固定為 100)。繪圖比較不同 Shrinkage 方法所計算出來的共變異數矩陣之 eigenvalues，同時也需與真實共變異數矩陣的 eigenvalues 及傳統共變異數矩陣的 eigenvalues 相比較 (參照講義 117/119, 119/119)。

6.45 Plot of Cook's Distance vs. Row Labels for airquality data。

- (a) 令 $y <- \text{airquality}\$Wind$, $x <- \text{airquality}\$Temp$ ，以此算出簡單線性迴歸之 MSE。
(Hint: $\text{MSE} = \sum(y_i - \hat{y}_i)^2 / (n - 2)$, 其中 n 為樣本個數)。
- (b) 以上述簡單線性迴歸模型為輸入，用 `plot` 只畫出 Cook's Distance vs. Row Labels 之二維圖。
- (c) 將原資料第一個資料點去除，重新 fit 簡單線性迴歸。印出其參數估計 (即 $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}$, 其中 $i = 1$)。
(Hint: $x[-1]$, $y[-1]$)。
- (d) 使用上題之參數估計，算出 fitted values 之平均。
(Hint: $\text{mean}, \hat{y}_{j(1)} = \hat{\beta}_{0(1)} + \hat{\beta}_{1(1)}x_j, j = 1, \dots, n$)
- (e) 算出第一點的 Cook's distance ($D_{(1)}$)。
(Hint: $D_{(i)} = \sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2 / (p \times \text{MSE})$, 其中 $p = 1, i = 1$ ，且 $\hat{y}_{j(i)}$ 為去掉第 i 點後之模型所估計出來的第 j 個 fitted values。)

(f) 算出所有點的 Cook's distance ($D_{(i)}$, $i = 1, \dots, n$) 之後，求其平均。

(Hint: `for(i in 1:n) .`)

(g) 畫出 Cook's Distance vs. Row Labels 之二維圖。

(Hint: `type="h" .`)

(h) 標出 Cook's distance 前三大值所在位置。

(Hint: `which, points, text .`)

6.46 Kernel density estimation (KDE, 核密度函數估計)

Let x_1, x_2, \dots, x_n be an iid sample drawn from some distribution with an unknown density f . We are interested in estimating the shape of this function f . Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

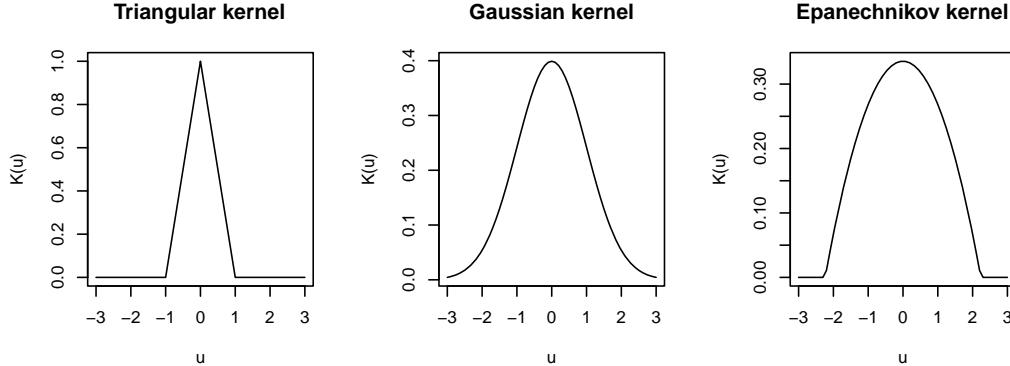
with kernel K and bandwidth h 。以下三個核函數 (kernel function) 是在進行核密度函數估計中常用的函數。

Kernel	Function
Triangular	$K(u) = (1 - u)I(u \leq 1)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Epanechnikov	$K(u) = \frac{3}{4\sqrt{5}}(1 - \frac{u^2}{5})I(u \leq \sqrt{5})$

$$\text{其中 } I(|u| \leq a) = \begin{cases} 1, & \text{if } |u| \leq a \\ 0, & \text{if } |u| > a \end{cases}$$

(a) 完成下列各題。

- 請將上述三種 kernel function 各寫成一 R 函式。提示: (1) `Triangular <- function(u) { ... }`, (2) `if`。
- 若 `u <- seq(-3, 3, 0.1)`，請畫出上述 kernel 圖形如下。



提示: (1) `apply(as.matrix(x) . . . ; (2) plot, par .`

- (b) 若觀察資料 x_1, x_2, \dots, x_n 為 $\text{xi} \leftarrow \text{iris[,1]}$ ，試寫一 R 函式，計算 $\hat{f}_h(x)$ 其在 $x = 7, h = 0.2736$ 之下，使用上述三種 kernel 之值。

提示:

```
> fh(xi, x=7, h=0.2736, kernel="Triangular")
[1] 0.1409978
> fh(xi, x=7, h=0.2736, kernel="Gaussian")
[1] 0.1797050
> fh(xi, x=7, h=0.2736, kernel="Epanechnikov")
[1] 0.1777105
```

- 6.47 一對夫婦計劃生孩子生到有女兒才停，或生了三個就停止。他們會擁有女兒的機率是多少？(印出電腦模擬 10 次的結果，及最後的機率。)

以電腦模擬計算機率的步驟如下：

第 1 步：機率模型每一個孩子是女孩的機率是 0.49，是男孩的機率是 0.51。各個孩子的性別是互相獨立的。

第 2 步：分配隨機數字。用兩個數字模擬一個孩子的性別: 00, 01, 02, …, 48 = 女孩; 49, 50, 51, …, 99 = 男孩

第 3 步：模擬生孩子策略。隨機產生一對一對的數字，直到這對夫婦有了女兒，或已有三個孩子。

第 4 步：計算機率。若 n 次重複中，有 m 次生女孩。會得到女孩的機率的估計是 m/n 。

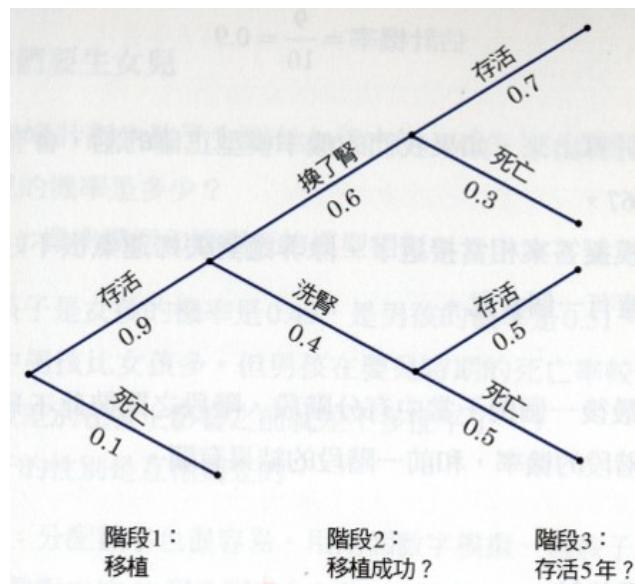
以下為模擬 10 次重複的範例，其中有 9 次生女孩，故得到女孩的機率的估計是 0.9:

69	05	16	48	17	87	17	40	95	17	84	53	40	64	89	87	20
男	女	女	女	男	女	女	男	女	男	男	女	男	男	男	女	
+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	+	

- 6.48 腎臟移植存活機率。(題目摘自「統計學的世界」一書)

腎臟移植的病人資料: 擁過移植手術的占 90%，另外 10% 會死亡。在手術後存活的人中有 60% 移植成功，另外的 40% 還是得回去洗腎。五年存活率對於換了腎的人來說是 70%，對於回去洗腎的人來說是 50%。計算能活過五年的機率。

- 第 1 步：機率模型如下圖。



- 第 2 步：對每個結果分配數字：
 - 階段 1: 0 = 死亡; 1, 2, 3, 4, 5, 6, 7, 8, 9 = 存活。
 - 階段 2: 0, 1, 2, 3, 4, 5 = 移植成功; 6, 7, 8, 9 = 仍需洗腎。
 - 階段 3: 換了腎: 0, 1, 2, 3, 4, 5, 6 = 存活五年; 7, 8, 9 = 未能存活五年。
 - 階段 3: 洗腎: 0, 1, 2, 3, 4 = 存活五年; 5, 6, 7, 8, 9 = 未能存活五年。(階段 3 的數字分配，和階段 2 的結果有關。所以二者間不獨立。)
- 例如: 在 4 次模擬中，有 2 次存活超過 5 年，則五年存活機率是 0.5。

	第1次	第2次	第3次	第4次
階段 1	3 存活	4 存活	8 存活	9 存活
階段 2	8 洗腎	8 洗腎	7 洗腎	1 新腎
階段 3	4 存活	4 存活	8 死亡	8 死亡

請寫一 R 程式 (命名為 `kidney_surgery_survival_prob`)，經 1000 次模擬，計算活過五年機率。

解: 本題之機率模型如上圖。令 X_1, X_2, X_3 代表各階段狀況的二元隨機變數: $X_i = 0, X_i = 1, i = 1, 2, 3$ 。令 $P(X_1 = 1) = 0.9$ ($P(X_1 = 0) = 0.1$) 為階段 1 病人手術後之存活(死亡)機率。令 $P(X_2 = 1|X_1 = 1) = 0.6$ ($P(X_2 = 0|X_1 = 1) = 0.4$) 為在階段 1 病人手術後存活狀況下，病人於階段 2 之移植成功(失敗後回去洗腎)的條件機率。令 $P(X_3 = 1|X_2 = 1, X_1 = 1) = 0.7$ ($P(X_3 = 1|X_2 = 0, X_1 = 1) = 0.4$) 為在階段 2 病人移植成功(失敗後回去洗腎)之下，病人於階段 3 之五年存活的條件機率。則病人能活過五年的機率為 $P(X_3 = 1) = P(X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_3 = 1|X_2 = 1, X_1 = 1) + P(X_1 = 1)P(X_2 = 0|X_1 = 1)P(X_3 = 1|X_2 = 0, X_1 = 1) = 0.9 * 0.6 * 0.7 + 0.9 * 0.4 * 0.5 = 0.558$ 。

6.49 (以 R 計算理論機率) 同上小題，將上述過程一般化: 腎臟移植的病人資料: 擰過移植手術的占 p ，另外 $(1 - p)$ 會死亡。在手術後存活的人中有 q 移植成功，

另外的 $(1 - q)$ 還是得回去洗腎。五年存活率對於換了腎的人來說是 s ，對於回去洗腎的人來說是 t 。計算能活過五年的機率。(其中 $0 < p, q, s, t < 1$,)。請寫一 R 函式，輸入為 (p, q, s, t) ，輸出為腎臟移植病人活過五年的(理論)機率。以 $(p = 0.9, q = 0.6, s = 0.7, t = 0.5)$ 試算機率。

7 資料處理

7.1 rbind {base} and cbind {base}

以下 exam1 及 exam2 為兩次小考成績。

```
set.seed(12345)
exam1 <- data.frame(student=paste0("A", sample(1:10)),
                      chinese=sample(0:100, 10, replace=T),
                      math=sample(0:100, 10, replace=T),
                      english=sample(0:100, 10, replace=T))

exam2 <- data.frame(student=paste0("A", sample(1:10)),
                      english=sample(0:100, 10, replace=T),
                      math=sample(0:100, 10, replace=T),
                      chinese=sample(0:100, 10, replace=T))
```

請將 2 次小考合併成下列格式:

	student	chinese	math	english	exam
10	A1	75	15	37	1
5	A10	39	8	59	1
3	A2	80	11	95	1
1	A3	93	83	79	1
9	A4	0	19	74	1
4	A5	31	2	97	1
8	A6	29	63	35	1
7	A7	37	12	24	1
2	A8	9	71	61	1
6	A9	38	13	31	1
51	A1	89	22	48	2
71	A10	98	25	55	2
81	A2	4	54	90	2
41	A3	11	54	66	2
101	A4	10	14	43	2
21	A5	45	85	12	2
61	A6	87	94	73	2
31	A7	86	50	94	2
91	A8	67	6	33	2
11	A9	100	34	82	2

7.2 (table {base})

record 為 500 位學生的年級和血型之紀錄 (含遺失值 NA)。

```

set.seed(123456)
n <- 500
grade <- as.factor(sample(c("大一", "大二", "大三", "大四"), n, replace=T))
bloodtype <- as.factor(sample(c("A", "AB", "B", "O"), n, replace=T))
bloodtype[sample(1:50, 30)] <- NA
record <- data.frame(grade, bloodtype)

```

請產生下列人數摘要統計表。

	bloodtype				
grade	A	AB	B	O	<NA>
大一	24	25	41	29	4
大二	34	29	27	34	8
大三	21	28	34	30	10
大四	28	32	27	27	8

7.3 (xtabs {stats})

Titanic 為 R 內鍵之資料集，儲存之類別為 array。請將此資料中的 Class 不計入 Crew 之人數印出如下。

```
, , Age = Child, Survived = No
```

Class	Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

```
, , Age = Adult, Survived = No
```

Class	Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	0	0

```
, , Age = Child, Survived = Yes
```

Class	Male	Female
1st	5	1
2nd	11	13
3rd	13	14
Crew	0	0

```
, , Age = Adult, Survived = Yes
```

Class	Male	Female
1st	57	140
2nd	14	80
3rd	75	76
Crew	0	0

7.4 (table {base})

資料檔 `Titanic_Kaggle_Dataset.xlsx` 為 Kaggle(<https://www.kaggle.com/c/titanic/data>) 提供的 Titanic 資料檔。變數資訊如下:

- Survived: Survival 0 = No, 1 = Yes
- Pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
- Sex
- Age: Age in years
- SibSp: number of siblings / spouses aboard the Titanic
- Parch: number of parents / children aboard the Titanic

- Ticket: Ticket number
- Fare: Passenger fare
- Cabin: Cabin number
- Embarked: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton 請將此資料讀入 R 中，並做以下轉換：

將 Age 大於 18 轉為 Adult，否則為 Child。將 Survived 轉為 "No"，"Yes"。將 Pclass 轉為 "1st"，"2nd"，"3rd"。最後建立 Pclass, Sex, Age 及 Survived 之資料陣列 (列聯表) 如下。

```
, , Age = Adult, Survived = No

Sex
Pclass female male
 1st      2   107
 2nd      6   131
 3rd     34   250

, , Age = Child, Survived = No

Sex
Pclass female male
 1st      1     4
 2nd      0    12
 3rd     21    61

, , Age = Adult, Survived = Yes

Sex
Pclass female male
 1st    118    36
 2nd     76     6
 3rd     60    27

, , Age = Child, Survived = Yes

Sex
Pclass female male
 1st     12     4
 2nd     21     9
 3rd     37    11
```

7.5 (xtabs {stats})

mobile.data 為一模擬信令資料，包含個人手機編號 id、日期時間 date.time、經度 longitude.x 及緯度 latitude.y。

```

set.seed(12345)
no.idv <- 5
id <- paste0("a", sprintf("%05d", 1:no.idv))
s <- strptime("02/27/2020 00:00:00", "%m/%d/%Y %H:%M:%S")
e <- strptime("02/27/2020 23:59:59", "%m/%d/%Y %H:%M:%S")
date.time <- rep(seq(s, e, by = "6 hour"), no.idv)
longitude.x <-
  sample(seq(120, 122, 0.01), 4 * no.idv, replace = T) #台灣東經 120度至 122度
latitude.y <-
  sample(seq(22, 25, 0.01), 4 * no.idv, replace = T) #台灣北緯 22度至 25度
mobile.data <- data.frame(id = rep(id, each = 4), date.time,
                           longitude.x, latitude.y)

```

請將 `mobile.data` 轉換成「經度」及「緯度」路徑資料如下。

	date.time				
id	2020-02-27 00:00:00	2020-02-27 06:00:00	2020-02-27 12:00:00	...	
a00001	121.41	120.50	121.51	...	
a00002	120.92	120.74	120.95	...	
a00003	120.85	120.74	120.37	...	
a00004	120.93	120.09	121.59	...	
a00005	121.66	120.37	120.29	...	

7.6 (stack {utils})

有一模擬資料 `simulated.data` 如下:

```

group <-rep(1:5, 4)
tag <-rep(LETTERS[1:4], each=5)
index <-matrix(rnorm(100), ncol=5)
simulated.data <-data.frame(group, tag, index=matrix(rnorm(100), ncol=5))

simulated.data
  group tag   index.1   index.2   index.3   index.4   index.5
1     1   A  1.0251185 -0.005172973  1.34425385  0.0128890660 -1.723029458
2     2   A  1.8713431  0.526850567 -0.38624966 -0.7781381624  0.330597394
3     3   A  0.5391122  0.958129585 -0.65496559 -1.0338181326  0.037590724
4     4   A -0.3485731  0.068696909  0.97496314 -1.8955495871 -1.123815765
5     5   A  0.6586629  0.103074268  0.38408186  0.4780770328  0.007415185
6     1   B -1.3355325 -0.909009407  0.77535950 -0.0004370019  1.285072776
7     2   B  0.3927303  1.084050657  1.58668619  0.3644630878  0.076876983
8     3   B  0.1198826  0.173335845  1.27255221 -0.4349315538 -0.448324691
9     4   B -0.0413866  0.277379166  1.75732829  1.1619234810  1.281319049
10    5   B  0.6252719 -1.601802622  0.14125203  0.0776981907  0.333032252
11    1   C  1.6602129  0.183001414 -0.08605801 -1.1112658054 -0.967540756
12    2   C  0.4628342  1.152702938 -0.68305502 -0.5459186538  1.134584660
13    3   C  0.3079730 -2.237561087  0.06286479 -0.4881918685 -1.348291919
14    4   C  0.1368502 -1.311245045 -0.62664304 -0.1356585838  1.930849990
15    5   C  0.3991671  1.738116258  1.04160411  1.2279369795  0.362777011
16    1   D  0.5581337 -0.992271175  0.63855315  0.5253661161 -0.698231197
17    2   D  1.1763206 -0.735594498  2.31932578 -0.7858969070  1.370117801
18    3   D -0.4946191 -0.835240634  0.04455638 -0.3234444673  0.765932329
19    4   D  1.6779458 -0.955329032  1.59524331  0.5931800483 -0.821282236
20    5   D -0.7125661 -1.250381114  0.81260201 -0.1544407470 -0.187501212

```

請轉成下列格式:

	group	tag	values	index
1		1	A 1.0251184510	index.1
2		2	A 1.8713431218	index.1
3		3	A 0.5391121730	index.1
4		4	A -0.3485730874	index.1
5		5	A 0.6586628783	index.1
6		1	B -1.3355325288	index.1
7		2	B 0.3927302522	index.1
8		3	B 0.1198826328	index.1
9		4	B -0.0413866004	index.1
10		5	B 0.6252718745	index.1
11		1	C 1.6602129233	index.1
		...		
92		2	C 1.1345846597	index.5
93		3	C -1.3482919194	index.5
94		4	C 1.9308499901	index.5
95		5	C 0.3627770110	index.5
96		1	D -0.6982311965	index.5
97		2	D 1.3701178007	index.5
98		3	D 0.7659323290	index.5
99		4	D -0.8212822357	index.5
100		5	D -0.1875012119	index.5

7.7 (unstack {utils})

chickwts 為 R 內鍵之資料集:

```
head(chickwts)
  weight feed
  1 179 horsebean
  2 160 horsebean
  3 136 horsebean
  4 227 horsebean
  5 217 horsebean
  6 168 horsebean
```

請轉成如下格式:

```

$casein
[1] 368 390 379 260 404 318 352 359 216 222 283 332

$horsebean
[1] 179 160 136 227 217 168 108 124 143 140

$linseed
[1] 309 229 181 141 260 203 148 169 213 257 244 271

$meatmeal
[1] 325 257 303 315 380 153 263 242 206 344 258

$soybean
[1] 243 230 248 327 329 250 193 271 316 267 199 171 158 248

$sunflower
[1] 423 340 392 339 341 226 320 295 334 322 297 318

```

7.8 (aggregate stats, tapply {base})

survey.data 為一模擬資料，包含變數 my.age、my.gender 及 my.score。

```

set.seed(12345)
my.age <- round(rnorm(100, mean=40, sd=20))
my.age <- ifelse(my.age < 0, 1, my.age)
my.gender <- sample(c("female", "male"), 100, replace=T, prob=c(0.6, 0.4))
my.score <- sample(0:100, 100, replace=T)
survey.data <- data.frame(id=paste0("a", 1:100), my.age, my.gender, my.score)

> head(survey.data)
  id my.age my.gender my.score
1 a1     52    female      65
2 a2     54      male      82
3 a3     38    female      71
4 a4     31    female      81
5 a5     52      male      36
6 a6      4      male      65

> tail(survey.data)
   id my.age my.gender my.score
95 a95     20      male      25
96 a96     54    female      40
97 a97     30      male      92
98 a98     83    female      43
99 a99     28    female      91
100 a100    26    female      98

```

請依性別與是否及格為分組，計算各組人數及各項變數平均如下。

```

pass
gender FALSE TRUE
female   35   23
male     23   19

```

```

gender pass my.age my.score
1 female FALSE 46.02857 27.45714
2 male   FALSE 46.69565 33.26087
3 female  TRUE 49.34783 78.60870
4 male   TRUE 36.15789 81.26316

```

7.9 (cut {base})

有一數字向量 normal.number 如下：

```

set.seed(12345)
normal.number <- rnorm(100, mean = 40, sd = 10)

```

請依數字範圍 (全距 range) 分成四組，並計算各組數字個數。

```

normal.number.cut1
(16.1,28.3] (28.3,40.5] (40.5,52.6] (52.6,64.8]
11           30           41           18

```

請依資料四分位數 (0% · 25% · 50% · 70% · 100%)，將資料分組，並計算各組數字個數。

```

normal.number.cut2
(16.2,34.1] (34.1,44.8] (44.8,49] (49,64.8]
24           25           25           25

```

7.10 (cut {base})

survey.data 為一模擬資料，包含 id, age, 和 gender。

```

set.seed(12345)
age <- round(rnorm(100, mean = 40, sd = 20))
age <- ifelse(age < 0, 1, age)
gender <- sample(c("female", "male"), 100, replace = T, prob = c(0.6, 0.4))
survey.data <- data.frame(id = paste("a", 1:100), age, gender)

```

請依 gender，將 age 分成 5 組 (全距等分方式)。

```
$female
[1] (40,60] (20,40] (20,40] (40,60] (20,40] (20,40] (20,40] (40,60]
... (中間省略)
[57] (20,40] (20,40]
Levels: (0,20] (20,40] (40,60] (60,80] (80,100]

$male
[1] (40,60] (40,60] (0,20] (20,40] (60,80] (20,40] (60,80] (0,20]
... (中間省略)
[41] (0,20] (20,40]
Levels: (0,20] (20,40] (40,60] (60,80] (80,100]
```

7.11 (replicate {base})

投擲兩顆公平的骰子，出現點數和為 $2, 3, \dots, 12$ 的理論機率值如下：

2	3	4	5	6	7	8	9	10	11	12
0.0278	0.0556	0.0833	0.1111	0.1389	0.1667	0.1389	0.1111	0.0833	0.0556	0.0278

試以 R 程式模擬投擲兩顆公平的骰子 (10000 次)，計算各點數和出現的機率。

2	3	4	5	6	7	8	9	10	11	12
0.0301	0.0592	0.0799	0.1122	0.1362	0.1570	0.1466	0.1157	0.0821	0.0527	0.0283

7.12 (lapply {base})

or sapply {base})

infert 為 R 內鍵之資料集，請判別各變數是否為數值型之變數。

education	age	parity	induced	case	spontaneous	stratum	pooled.stratum
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

7.13 (mapply {base})

anscombe 為 R 內鍵之資料集：

x1	x2	x3	x4	y1	y2	y3	y4	
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04

請計算下列兩兩變數之相關係數。

cor(x1, y1)	cor(x2, y2)	cor(x3, y3)	cor(x4, y4)
0.8164205	0.8162365	0.8162867	0.8165214

8 資料分析

8.1 「癌症發生統計」之資料摘要：資料來源：政府資料開放平台。資料檔：「File_15197.csv」。
<https://data.gov.tw/dataset/6399>

- (a) 讀取資料，印出資料前後各 10 筆紀錄。
- (b) 以 ggplot2，畫出台北市「癌症別：肺、支氣管及氣管」歷年「癌症發生數」之長條圖。
- (c) 依「縣市別」及「性別（不分性別、男。女）」之分類，癌症發生總數各是多少人？
- (d) 依「性別（不分性別、男。女）」，列出癌症發生總數最高之前 5 名的癌症別及其總數。（例：男性之下，各癌症別發生總數 = 加總所有「癌症診斷年」及「縣市別」之下的各癌症別發生數）

8.2 (探索性資料分析)

ChickWeight {datasets} 是 R 內建的資料。紀錄小雞在不同飼料餵食之下的體重。關於此資料更多說明，請「?ChickWeight」。利用所學之統計圖探索此資料。（畫出之圖不需解說，但圖之標題，xy 軸標號及圖例說明需完整，讓讀者一看就可得知資訊；以最多資料呈現在最少張圖上為原則）（提示：summary）

8.3 (探索性資料分析)

資料來源：政府資料開放平台。資料檔：不動產實價登錄資訊-買賣案件。<https://data.gov.tw/dataset/26820>

- (a) 下載資料，並讀入 RStudio。印出資料摘要（summary）及結構（str）。請確認每一變數（欄位）皆是正確的 R 類別（例如：數值變數、日期變數，文字的「一、二、三...」等等轉成數值型的「1,2,3,...」）。若不是請做必要的類別轉換。
- (b) 呈上題之「不動產實價登錄資訊-買賣案件」資料。依 district(鄉鎮市區)，計算此資料各個連續變數之平均。例如：依各 district(鄉鎮市區)，變數「rps22(單價每平方公尺)」之平均為何。
- (c) 進行探索性資料分析。（畫出各式統計圖（包含索引圖、直方圖、長條圖、盒形圖、餅圖、2D 散佈圖、3D 散佈圖、熱圖、地圖等等）（需對圖形結果作一些簡單解釋。）（請自行編子題號：2.1, 2.2, ...）
- (d) 依此資料探索之結果，列舉一些想了解或解決的問題。這些問題需要其它的輔助資料嗎？若有請說明。

8.4 (探索性資料分析)

資料集來源：政府資料開放平臺：「家庭收支調查-家庭消費支出結構按消費型態分」

- 資料概述：<https://data.gov.tw/dataset/6588>

- 請於資料網頁「資料資源」一項中，選按「檢視資料」，下載「CSV」檔:
”48309de5de430725c28d9855fd3f7af4_export.csv”
- 對此資料做探索性資料分析 (EDA)。(繪圖時，變數 Year 請採用地形色階呈現)
 - 試以 Principal Components Analysis 分析此資料。需解釋或說明所列出的報表、統計量、圖形等等代表的意義及現象。(提示: (1) 是否需標準化? (2) 繪圖時，變數 Year 請採用地形色階呈現)

8.5 (探索性資料分析)

- 資料名稱: 小麥種子 (seeds Data Set)
 - 資料來源: <https://archive.ics.uci.edu/ml/datasets/seeds>
 - 資料說明: 有三個不同品種的小麥種子: Kama, Rosa 和 Canadian，每一品種小麥皆隨機選擇出 70 顆麥粒，接受檢測。每顆麥粒量測以下 7 種幾何屬性: (1) 面積 (area, A); (2) 周長 (perimeter, P); (3) 緊密度 (compactness $C = 4\pi A/P^2$); (4) 麥粒長度 (length of kernel); (5) 麥粒寬度 (width of kernel); (6) 不對稱係數 (asymmetry coefficient); (7) 麥核溝長度 (length of kernel groove)。
- 讀取資料，印出資料摘要。
 - 利用所學之統計圖探索此資料。(畫出之圖不需解說，但圖之標題， xy 軸標號及圖例說明需完整，讓讀者一看就可得知資訊；以最多資料呈現在最少張圖上為原則) (提示: summary)
 - 以 plot3D 套件，畫出 area, perimeter 和 asymmetry coefficient 的 3D 散佈圖，其中，圖上的點以紅、綠、藍三個顏色代表三個不同品種的小麥種子。
 - 將每個屬性變數做標準化 $z = (x - \bar{x})/s$ 。印出標準化資料摘要。
 - 利用 heatmap 指令，畫出此標準化後資料的熱圖 (heatmap)。(註: 使用內建距離量測尺度及群集分析方法即可。熱圖旁需有小麥品種的顏色條)

8.6 (探索性資料分析)

資料概述:

- 行政院環境保護署空氣品質監測網 <http://taqm.epa.gov.tw/taqm/tw/YearlyDataDownload.aspx>
- 北部空品區 105 年監測資料檔 (105_HOUR_01_20170301.zip) (註: 此監測網資料包含全台灣地區，因時間關係，僅取北部空品區練習)
- 空氣品質監測 105 年年報: 105_YEAR_00.pdf ([重要] 105 年年報已有完整的問題、分析及圖表，你可以參考裡面一些背景知識，發掘一些問題，也可以使用 R 將它裡面的圖表重覆再做一次，驗証看看。)

請運用目前課程所學，探索「空氣品質」資料:

- 讀取全部檔案，並利用一些圖形或統計量檢查資料之正確性。(資料有沒有問題？你可能要知道每一變數之觀察值合理的範圍是什麼。)

- (b) 資料的基本統計量為何？觀察值的範圍為何？分佈為何？
- (c) 條列出你可能想了解的問題（列出問題即可，不要管能不能解決或不切實際；例如：群組比較、變數間的相關或影響等等）。
- (d) 以上的問題，我期待（猜測）的答案（結果）是什麼？需要什麼額外的資料來輔助分析嗎？（可能經過分析之後，答案不一定是正確的，沒關係）
- (e) 空氣品質指標（AQI）的定義：<http://taqm.epa.gov.tw/taqm/tw/b0201.aspx> 維基百科、搜尋「空氣品質指數」，查詢公式。<https://zh.wikipedia.org> 為簡化起見，假設「污染物項目濃度」在計算 AQI 過程中是以 24 小時平均值為標準。請計算「板橋」站在 105 年度 12 個月份之空氣品質指標（AQI）。
- (f) 將資料整理成以下兩個 data.frame，（取名 airdata.mean，airdata.var），表格中的 value 是一整個月（一天有 24 個紀錄值）的平均數（遺失值不列入計算）或變異數。格式如下：

Area	Year	Month	AMB.TEMP	CH4	...
三重站	2016	1	value		
		2			
		⋮			
		12			
土城站	2016	1			
⋮	⋮	2			
⋮	⋮	⋮			
⋮	⋮	12			
⋮	⋮	⋮			
觀音站		1			
⋮	2016	12			

- (g) 利用 airdata.mean（airdata.var）畫出每一監測站 PM2.5 之時間序列圖：橫軸為（1~12 月），縱軸是 PM2.5 月平均值（變異數）。圖中每一條線代表一個監測站。兩張圖你有什麼發現？
- (h) 以 airdata.mean 為例，CO，SO2 兩污染物（變數）的分佈為何？請做 QQplot 及常態分佈檢定（參考老師講義）。需要考慮做資料轉換嗎？試著使用三種不同資料轉換（其中一個是 Cox-Box），並解釋為何要採用所選的轉換方式。轉換前後有什麼差別？
- (i) 依照「B01-2：遺失值、離群值處理，76/84」之準則，自選 4 種遺失值補值方法，評估哪一個是最佳的。

```

tmp <- airdata.mean[, -(1:3)]
np <- nrow(tmp) * ncol(tmp)
id <- sample(1:np, floor(np* 0.1))
tmp[id] <- NA
airdata.mean.miss <- cbind(airdata.mean[,1:3], tmp)

```

- (j) 答案卷最後可列出參考的網站、書本、或參考資料。

8.7 (探索性資料分析)

資料檔：「薪情平臺匯出資料.xlsx」。資料來源：薪資平台查詢系統 (https://earnings.dgbas.gov.tw/query_payroll.aspx)

- (a) 直接讀入資料檔「薪情平臺匯出資料.xlsx」，產生兩個資料框 (data.frame) 「每人每月總薪資 (新臺幣元)」及「總工時 (小時)」。其中資料框中每一欄位為每一行業別，並新增一欄位「性別」。
- (b) 印出資料摘要 (summary) 及結構 (str)。請確認每一變數 (欄位) 皆是正確的 R 類別 (例如：數值變數、日期變數)。若不是請做必要的轉換。
- (c) 針對「每人每月總薪資 (新臺幣元) · 男性」，畫出時間序列圖：橫軸為時間，縱軸為薪資，圖上每一條趨勢線代表每一行業別之薪資。需標出圖例說明。需對結果作一些簡單解釋。
- (d) 畫出 side-by-side 盒形圖：橫軸為每一行業別 (請依照每一行業別之中位數從大至小，由左至右排序)，縱軸為近五年 (102~106) 薪資之五數綜合。需對結果作一些簡單解釋。
- (e) 畫出兩個資料框之熱圖 (heatmap)，其中需有「性別」之色條。(可搭配群集分析對產業別排序) 需對結果作一些簡單解釋。
- (f) 針對此資料「每人每月總薪資 (新臺幣元)」及「總工時 (小時)」，自行問一個想了解的問題，並進行探索性資料分析。需對結果作一些簡單解釋。

8.8 (探索性資料分析)

資料來源：政府資料開放平台。資料檔：「癌症發生統計.csv」。

<https://data.gov.tw/dataset/6399>

- (a) 請對此資料進行探索性資料分析。
(應至少包含：列印資料前幾筆紀錄，各 (連續及類別) 變數之統計圖表、敘述統計，及對這些結果之簡單說明) (可自行編小題號)
- (b) 此資料中，依「縣市別」及「性別 (不分性別、男。女)」之分類，癌症發生總數各是多少人？
- (c) 此資料中，依「性別 (不分性別、男。女)」，列出癌症發生總數最高之前 5 名的癌症別及其總數。(例：男性之下，各癌症別發生總數 = 加總所有「癌症診斷年」及「縣市別」之下的各癌症別發生數)
- (d) 以「胃癌」為例，依「性別 (不分性別、男。女)」之類別，計算每年「年齡中位數」之平均數。

8.9 (探索性資料分析)

資料檔 SalaryGov_Month.xlsx 為政府薪情平臺匯出之「每人每月總薪資 (新臺幣元)」資料 (時間為 101 年 8 月至 107 年 8 月)。其中每一「行業類別」下，含有「性別」欄位。

- (a) 讀入資料檔，印出資料摘要 (summary) 及結構 (str)。請確認每一變數 (欄位) 皆是正確的 R 類別 (例如：數值變數、日期變數)。若不是請做必要的轉換。

- (b) 針對「每人每月總薪資 (新臺幣元) · 男性」，畫出時間序列線圖：橫軸為時間，縱軸為薪資，圖上每一條趨勢線代表每一行業別之薪資。需標出圖例說明。需對結果作一些簡單解釋。
- (c) 畫出 side-by-side 盒形圖：橫軸為每一行業別（請依照每一行業別之中位數從大至小，由左至右排序），縱軸為近五年（101/08~107/08）薪資之五數綜合。需對結果作一些簡單解釋。
- (d) 畫出資料之熱圖（heatmap），其中欄位需有「性別」之色條，列位需有「民國年」之色條。（可搭配群集分析對產業別排序）需對結果作一些簡單解釋。（用 pheatmap 套件畫）

8.10 (探索性資料分析 +CCA)

資料集來源：政府資料開放平臺：「用電統計資料」

- 資料概述：<https://data.gov.tw/dataset/6064>
- 請於上述網頁下載資料檔（或附檔 electricity-001.csv）：

（壓縮檔，內含「歷年平均單價.txt、歷年用戶數.txt、歷年行業別（本檔不再更新）.txt」）

 - (a) 對此資料做探索性資料分析（EDA）。(註：繪圖時，變數”年別（民國）/民國年”請採用彩虹色階呈現 (tim.colors {fields}))
 - (b) 資料處理（有需要做資料處理嗎？例如標準化、轉換、刪除某些觀察值、選取某些變數進行分析等等？）
 - (c) 試以 Canonical Correlation Analysis 分析此資料。需解釋或說明所列出的報表、統計量、圖形等等代表的意義及現象。(註：繪圖時，變數”年別（民國）/民國年”請採用彩虹色階呈現 (tim.colors {fields}))

8.11 (敘述統計)

資料來源：政府資料開放平台。資料檔：「癌症發生統計.csv」。<https://data.gov.tw/dataset/6399>

- (a) 以「癌症別：肝及肝內膽管」為例，依「性別（不分性別、男。女）」之類別，計算「年齡標準化發生率」之平均數、中位數及眾數。
- (b) 以「癌症別：胃」為例，畫出各縣市別的「平均癌症發生數」的長條圖。（各縣市別的平均癌症發生數的算法：各縣市別的歷年（癌症診斷年）的癌症發生數總和除以診斷年個數。）

8.12 (常態分配檢定)

小明想知道資料 data(swiss) 中的 Fertility 變數是否來自常態分配，請你用 R 幫他分析一下。（提示：QQplot 及常態分佈檢定法：ks.test, ad.test, shapiro.test）

8.13 (假設檢定)

資料來源: 政府資料開放平台。資料檔: 「消費者端量測行動上網平均速率.csv」。

資料網址: <https://data.gov.tw/dataset/8258>

- 畫出資料中第一階段、第二階段的盒形圖。(提示: side-by-side boxplot)
- 請問消費者端量測行動上網平均速率第一階段與第二階段是否有顯著差異?
(提示: (1) 有母數及無母數方法，皆各選一用合適的檢定。(2) 有母數方法需注意資料是否符合假設)

8.14 (變異數分析/事後檢定)

資料來源: 政府資料開放平台。資料名稱: 「各國證券市場成交值週轉率比較 NEW」。資料檔: 「每月_103938_A43_t35 世界主要證券市場成交值周轉率比較(35).csv」。資料網址: <https://data.gov.tw/dataset/103938>

- 選取 2001/01~2002/12 之資料(除「上海」外)，儲存成一資料框(data.frame)，並列印出。
- (承上小題) 畫出每個地區(臺灣, 紐約, 日本, 倫敦, 香港, 韓國, 新加坡)之盒形圖。(提示: side-by-side boxplot)
- (承上小題) 每個地區(臺灣, 紐約, 日本, 倫敦, 香港, 韓國, 新加坡)之成交值週轉率是否有顯著差異？若有差異，是哪些地區有差異？

8.15 (遺失值處理)

- 資料來源: UCI Machine Learning Repository
- 資料集: Mammographic Mass Data Set
- 網址: <https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>
- 資料檔: mammographic_masses.data
- 說明檔: mammographic_masses.names

- 如何觀察此資料遺失值之樣態？(請實作)
- 哪一種補值方法較好？(至少試 3 種(含)以上的補值方式)

8.16 (遺失值處理)

以下為模擬具有遺失值資料 x 之 R 程式碼:

```
n <- 1000
p <- 10
set.seed(123456)
library(MASS)
s <- matrix(rt(p*p, df=5), ncol = p)
sigma <- crossprod(s)
x <- mvrnorm(n, mu=rep(0, p), Sigma=sigma)
missing.percentage <- 0.1
x[sample(n*p, floor(n*p*missing.percentage))] <- NA
```

- (a) 選取完整之資料 (命名為 `x.complete`)，印出此資料之維度 ($nc \times pc$)。
- (b) 模擬遺失：將上述之資料隨機選取出比例為 `missing.percentage` 之觀察值 (ξ_i)，設置成 NA(命名 `x.complete.na`)。
 (提示: `set.seed(54321); ij <- sample(1:(nc*pc), floor(nc*pc*missing.percentage))`)
- (c) 利用下列 5 方法各自對上述資料 (`x.complete.na`) 做補值：Mean Substitution, K-Nearest Neighbour Imputation (K=5), `mice.impute.pmm` {MICE}, `mice.impute.norm` {MICE}。
- (d) 計算下列指標數值，評估上述 5 種補值方法：

$$\sum_{i=1}^m (\hat{\xi}_i - \xi_i)^2,$$

其中 $m=floor(nc*pc*missing.percentage)$ 、 ξ_i 為模擬遺失之真實值， $\hat{\xi}_i$ 為 ξ_i 之補值。

8.17 (資料轉換)

資料來源: (UCI) Concrete Compressive Strength Data,
<http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>。
 說明檔見: 「Concrete_Readme.txt」

- (a) 讀取資料 `Concrete_1030x9.txt`，並做多重迴歸分析 (`lm`)，其中 y 為反應變數，{Cement, BFS, FlyA, Water, Sp, CA, FineA, Age} 為解釋變數，印出 R^2 值。
- (b) 對資料做 (至少 5 種方法) 轉換 (部份或全部的解釋變數)，(方法其中至少包含標準化及 Box-Cox 轉換)，(可以有複合式轉換，例如標準化後，再施行另一種轉換)，並將轉換後的資料以多重迴歸方法分析，印出 R^2 值。對此資料而言，那一種轉換可以得到較高的 R^2 值？

8.18 (資料轉換)

- 資料來源: UCI Machine Learning Repository
- 資料集: Airfoil Self-Noise Data Set
- 網址: <https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise>
- 資料檔: `airfoil_self_noise.dat`

此資料之目的是建立一迴歸模型，利用 5 個解釋變數 (Frequency, Angle of attack, Chord length, Free-stream velocity, Suction side displacement thickness) 來預測一反應變數 (Scaled sound pressure level)。此資料有需要做轉換嗎？若有需要，請至少試 3 種 (含) 以上的資料轉換方式，並比較哪一個資料轉換方式較適合迴歸模型。

8.19 (簡單線性迴歸模型)

選取 `iris` 資料之 `Petal.Length` 為解釋變數，`Petal.Width` 為反應變數。

- (a) 畫出此兩變數之散佈圖 (圖 A)，並加上 Species 的顏色。
- (b) 配適此兩變數之簡單線性迴歸模型：列出配適結果及 anova 表格。
- (c) 在散佈圖 A 上加上迴歸線並在圖上註明其迴歸方程式。
- (d) 畫出 Fitted values 和 Residuals 之散佈圖 (圖 B)，並加上一水平方程式為 0 的線。
- (e) 在圖 B 上標出 Residuals 最大及最小的值。
- (f) 去掉最後 50 個觀察點，利用 update，重新配適一簡單線性迴歸模型，並列出配適結果。
- (g) 利用以上兩個模型及 predict 指令，計算當 Petal.Length 為 2.5 及 3.2 時，Petal.Width 的預測值。

8.20 (迴歸分析)

資料來源: (UCI) Concrete Compressive Strength Data,
<http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>。
說明檔見: 「Concrete_Readme.txt」

- (a) 讀取資料 Concrete_1030x9.txt，並做多重迴歸分析 (lm)，其中 y 為反應變數，{Cement, BFS, FlyA, Water, Sp, CA, FineA, Age} 為解釋變數，印出 ANOVA 表，並解釋 R^2 值及 F 值。(注意：解釋變數需做標準化)
- (b) 請以 R 內建之六個統計模型檢測圖形，說明上小題之迴歸模型之假設是否合適？若不合適，資料變數應該做哪些轉換 (例如: \log , $1/y$ 等等，請見一般迴歸分析之教科書)？資料中是否有離群值 (outliers or leverages)？
- (c) 進行逐步迴歸變數篩選。

8.21 (群集分析)

- 資料來源: UCI Machine Learning Repository。
 - 資料名稱: 「QSAR aquatic toxicity Data Set」。
 - 資料檔: 「qsar_aquatic_toxicity.csv」。
 - 資料網址: <https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity>
 - 變數資訊: 此資料共 546 觀察值。變數依序為「TPSA(Tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, C-040, LC50」，其中反應變數為「LC50」。
- (a) 讀取資料，印出資料前 5 筆及後 5 筆紀錄。
 - (b) 使用 pheatmap 套件對 8 個解釋變數及 546 觀察值 (化學物質) 進行階層式群集分析 (Complete-linkage + Euclidean distance)，並畫出熱圖。其中 1 個反應變數是註解，以顏色條呈現於熱圖旁。
 - (c) 使用 clValid 套件，針對 546 觀察值進行群集驗証 (內在指標與穩定性指標)，比較“kmeans”，“diana”，“fanny” 及“pam” 四種群集方法，群數設為 3~7 群，其它參數設定為預設值，印出各方法的指標比較報表及指標 vs 群數的線圖。

8.22 (迴歸分析)

- 資料來源: UCI Machine Learning Repository。
- 資料名稱: 「QSAR aquatic toxicity Data Set」。
- 資料檔: 「qsar_aquatic_toxicity.csv」。
- 資料網址: <https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity>
- 變數資訊: 此資料共 546 觀察值。變數依序為「TPSA(Tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, C-040, LC50」，其中反應變數為「LC50」。
 - (a) 進行迴歸分析，並印出參數估計報表及 ANOVA 表格，需做簡單解釋。
 - (b) 進行逐步迴歸分析，選取最佳之模型。

8.23 (迴歸分析、群集分析)

- 資料來源: UCI Machine Learning Repository。
- 資料名稱: 「QSAR aquatic toxicity Data Set」。
- 資料檔: 「QsarAquaticToxicity_std.csv」。(已附上標準化之資料檔)
- 資料網址: <https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity>
- 變數資訊: 此資料共 546 觀察值 (化學物質)。8 個解釋變數 (分子描述)，1 個反應變數 (水蚤的定量水生急性毒性)，依序為「TPSA(Tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, C-040, LC50」，其中反應變數為「LC50」。
 - (a) 讀取資料，印出資料前 5 筆及後 5 筆紀錄。
 - (b) 畫出所有變數的散佈圖矩陣。
 - (c) 畫出所有變數的索引圖 (使用 ggplot2，並以一頁 3×3 的圖呈現)。
 - (d) 進行多重迴歸分析，印出參數估計報表及 ANOVA 表格。
 - (e) (承上題) 預測反應變數 LC50 之估計值，當 (TPSA.Tot, SAacc, H.050, MLOGP, RDCHI, GATS1p, nN, C.040) 之值為 $(-1.2765, -0.0266, -0.3740, 1.8024, 1.1817, -2.3618, -0.3548, 0.4596)$ 。
 - (f) 進行逐步迴歸分析，選取最佳之模型。
 - (g) 使用 8 個解釋變數，對 546 觀察值 (化學物質) 進行 K-均值法分群，設定 $K=3, 4, 5, 6$ 時，印出各自的群內平方總和 (total within-cluster sum of square)。

8.24 (維度縮減)

資料集來源: Wine Data Set, <https://archive.ics.uci.edu/ml/datasets/wine>
 以下 ISOMAP 演算法中，鄰居個數一律設定為 5，若自覺得不合適，請自行選一“合適”的個數。

- (a) 讀取資料，以 MDS 及 ISOMAP 做維度縮減，各得到前兩維的維度縮減資料，畫出散佈圖，其中圖上的點 (酒) 若為同一品種，則以同顏色顯示。(酒彼此之間的距離為歐式空間距離)

- (b) 使用 ISOMAP 做維度縮減，以 rgl 套件畫出前三維之 3D 散佈圖。(點的顏色顯示要求同上，3D 散佈圖請轉三個不同角度貼上答案卷)
- (c) 於上小題之 3D 散佈圖中，若是兩點鄰居，則加一連線。(點的顏色顯示要求同上，3D 散佈圖請轉三個不同角度貼上答案卷)
- (d) 將資料隨機分成 2/3 訓練集及 1/3 測試集，利用訓練集造出線性 SVM 模型，印出測試集的分類 cross table，並算出測試集的分類錯誤率。(set.seed(12345))
- (e) 使用 ISOMAP 做維度縮減得到的投影資料記做 $Z_{n \times k}$ ，其中 n 為酒的個數， $k = 1, 2, \dots, 10$ 為所取的低維度個數。當 $k = 1, 2, \dots, 10$ 時，將資料 $Z_{n \times k}$ 隨機分成 2/3 訓練集及 1/3 測試集，利用訓練集造出線性 SVM 模型，計算出測試集的分類錯誤率。(set.seed(12345))
- (f) 使用 MDS，重覆上小題之步驟。
- (g) 將前 3 小題的三種方法所得到的錯誤率，繪出線圖。橫軸為 $k = 1, 2, \dots, 10$ ，縱軸為錯誤率。(需加上方法的圖例說明 (legend)。)
- (h) 算出 Wine Data Set 的 geodesic distance (即 IsoDistance)，記做 D_G ，使用 hclust {stats} 在 D_G 上做群集分析，並使用 cutree {stats} 將 hclust 結果分成 3 群，將分群結果以顏色呈現在 ISOMAP 的前兩維散佈圖上。
- (i) 以 LCMC 及 STRESS 評估 MDS 及 ISOMAP 維度縮減的表現。 $(d_{ij}, \hat{d}_{ij}$ 請參照講義)

$$STRESS = \sum_{i < j}^n (d_{ij} - \hat{d}_{ij})^2.$$

8.25 (維度縮減)

資料來源: (UCI) Statlog (Vehicle Silhouettes),

[https://archive.ics.uci.edu/ml/datasets/Statlog+\(Vehicle+Silhouettes\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes))，其中 class 為具有 4 個類別的反應變數 (Y)，{compactness, circularity, …, hollows} 為 18 個解釋變數 (X)。

- (a) 讀取資料 statlog_vehicle_846x18.txt，利用下列 4 種維度縮減法針對 X 做維度縮減: PCA · MDS · ISOMAP · SIR。各畫出其降維後的二維散佈圖，每一觀察值需以顏色標上類別。(一頁 4 張圖)
- (b) 承上小題，各畫出 circle of correlations 圖。(一頁 4 張圖)
- (c) 承上小題，畫出 PCA 特徵值之陡坡圖。

8.26 (維度縮減)

資料來源 (UCI): <https://archive.ics.uci.edu/ml/datasets/glass+identification>
Glass Identification Data Set 是玻璃識別資料集，共有 214 個觀察值，具有 9 個變數 (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) 及一個類別變數 (class · 6 個類別)。

- (a) 使用 PCA {FactoMineR} 對此資料進行主成份分析。((1) 印出 Eigenvalues/Variances，畫出 scree plot。(2) 畫出 Circles of Correlation 圖，圖上各點以

square cosine (cos2 values) 之值為顏色標示出來。(3) 畫出 214 觀察值投影於前兩個主成份之散佈圖，並以不同顏色標示 6 個類別。)

- (b) 對此資料進行 MDS 及 ISOMAP 維度縮減方法，並畫出維度縮減後的資料於二維平面的投影散佈圖，同時以不同顏色標示 6 個類別 (請自行選用合適之輸入參數)

8.27 (共變異數矩陣 + 維度縮減)

以下三個 R 套件皆提供一些函數可計算 Shrinkage estimation of covariance matrix:

- cov.shrink {corpcor}
- shrinkcovmat.identity {ShrinkCovMat}
- covEstimation {RiskPortfolios} with type = 'oneparm'

現以 R 程式產生一模擬資料 x 如下：

```
library(MASS)
n <- 10
p <- 100
set.seed(123456)
sigma <- matrix(rnorm(p * p), ncol = p)
sigma <- crossprod(sigma) + diag(rep(0.2, p))
x <- mvrnorm(n, mu=rep(0, p), Sigma=sigma)
```

試以不同的 p/n 值 ($p/n = 0.1, 0.5, 2, 10$, p 固定為 100)，繪圖 (一頁 4 張圖) 比較不同 Shrinkage 方法所計算出來的共變異數矩陣之 eigenvalues，同時也需與真實共變異數矩陣的 eigenvalues 及傳統共變異數矩陣的 eigenvalues 相比較 (參照 DR 講義 117-119/144)(一張圖上共 5 條線，每一條線代表一個共變異數矩陣計算方法)。

- 8.28 承上題之 4 組模擬資料 ($p/n = 0.1, 0.5, 2, 10$, $p = 100$) (各命名為 x.bn0.1, x.bn0.5, x.bn2, x.bn10) · 以主成份分析 (PCA) 及等軸距特徵映射 (ISOMAP) 做維度縮減 (維度為 2~5) · 計算 LCMC 評估指標 (其中鄰居個數 (neighborhood size) 範圍為 5~10)。 (使用之方法參數，請自行選合適的。PCA/ISOMAP 請說明所使用的套件或修正方法) · 將結果以下列格式列出：

```

x.pn0.1
PCA
dim   K=5   K=6   K=7   K=8   K=9   K=10
2     ...   ...   ...   ...   ...   ...
3     ...   ...   ...   ...   ...   ...
4     ...   ...   ...   ...   ...   ...
5     ...   ...   ...   ...   ...   ...

ISOMAP
dim   K=5   K=6   K=7   K=8   K=9   K=10
2     ...   ...   ...   ...   ...   ...
3     ...   ...   ...   ...   ...   ...
4     ...   ...   ...   ...   ...   ...
5     ...   ...   ...   ...   ...   ...

x.pn0.5
PCA
dim   K=5   K=6   K=7   K=8   K=9   K=10
2     ...   ...   ...   ...   ...   ...
3     ...   ...   ...   ...   ...   ...
4     ...   ...   ...   ...   ...   ...
5     ...   ...   ...   ...   ...   ...

ISOMAP
dim   K=5   K=6   K=7   K=8   K=9   K=10
2     ...   ...   ...   ...   ...   ...
3     ...   ...   ...   ...   ...   ...
4     ...   ...   ...   ...   ...   ...
5     ...   ...   ...   ...   ...   ...

...

```

8.29 (典型相關分析, Canonical correlation analysis) 某商業公司為了瞭解會影響旗下業務員銷售表現的因子，對 50 位業務員做了一個調查，搜集以下兩組變數：銷售表現 (Sales Performance: Sales Growth、Sales Profitability、New Account Sales) 及智力測驗分數 (Intelligence Scores: Creativity、Mechanical Reasoning、Abstract Reasoning、Mathematics)。請對此兩組變數做典型相關分析。(提示: (1) 資料集: sales-cca.csv。(2) 讀入資料後，請依變數名重新指定欄位名稱。)

8.30 正規化典型相關分析 (Regularized CCA)

- 資料集名稱: Lung Cancer Microarray Data (73x831)

- 資料集檔案: LungCancer_Garber2001_73sx918g.txt
- 資料集描述: 維度縮減講義: 第 107/144 頁

請將 genes 以 kmeans 分成兩群, 並針對此兩群 genes(變數) 進行正規化典型相關分析。

8.31 (群集分析)

資料集: Morphological Measurements on Leptograpsus Crabs: The crabs data frame has 200 rows and 8 columns, describing 5 morphological measurements on 50 crabs each of two colour forms and both sexes, of the species Leptograpsus variegatus (紫岩蟹) collected at Fremantle, W. Australia.

```
library(MASS)
data(crabs)
names(crabs)
?crabs
crabs2 <- crabs
crabs2$index <- rep(1:4, each=50)
```

- 以 pheatmap 套件畫出資料 crabs2 ("FL", "RW", "CL", "CW", "BD") 未分群前之熱圖 (heatmap)。(熱圖邊需加上"sp", "sex", "index" 之資訊，且需有圖例說明 (legend))
- 對此資料 (200 隻蟹) 做 K 均值法群集分析 ($K = 4$)。5 個欄位變數則做 average-linkage，將分群結果，一同呈現在上述的 heatmap 中。(即新增一彩色邊條為 K 均值法之結果，整個資料熱圖之排序，依序是 K 均值法結果、"sp"、"sex"、"index"。)(資料排序提示: arrange {dplyr})
- (承 (a) 小題) 以 pheatmap 套件畫出資料 crabs2 (蟹及變數) 之 complete-linkage 和 single-linkage 之熱圖。(熱圖邊需加上"sp", "sex", "index" 之資訊，且需有圖例說明 (legend))(需將兩熱圖畫在一起比較，即一頁兩張圖) (熱圖請使用與 (a) 小題不同之色階)

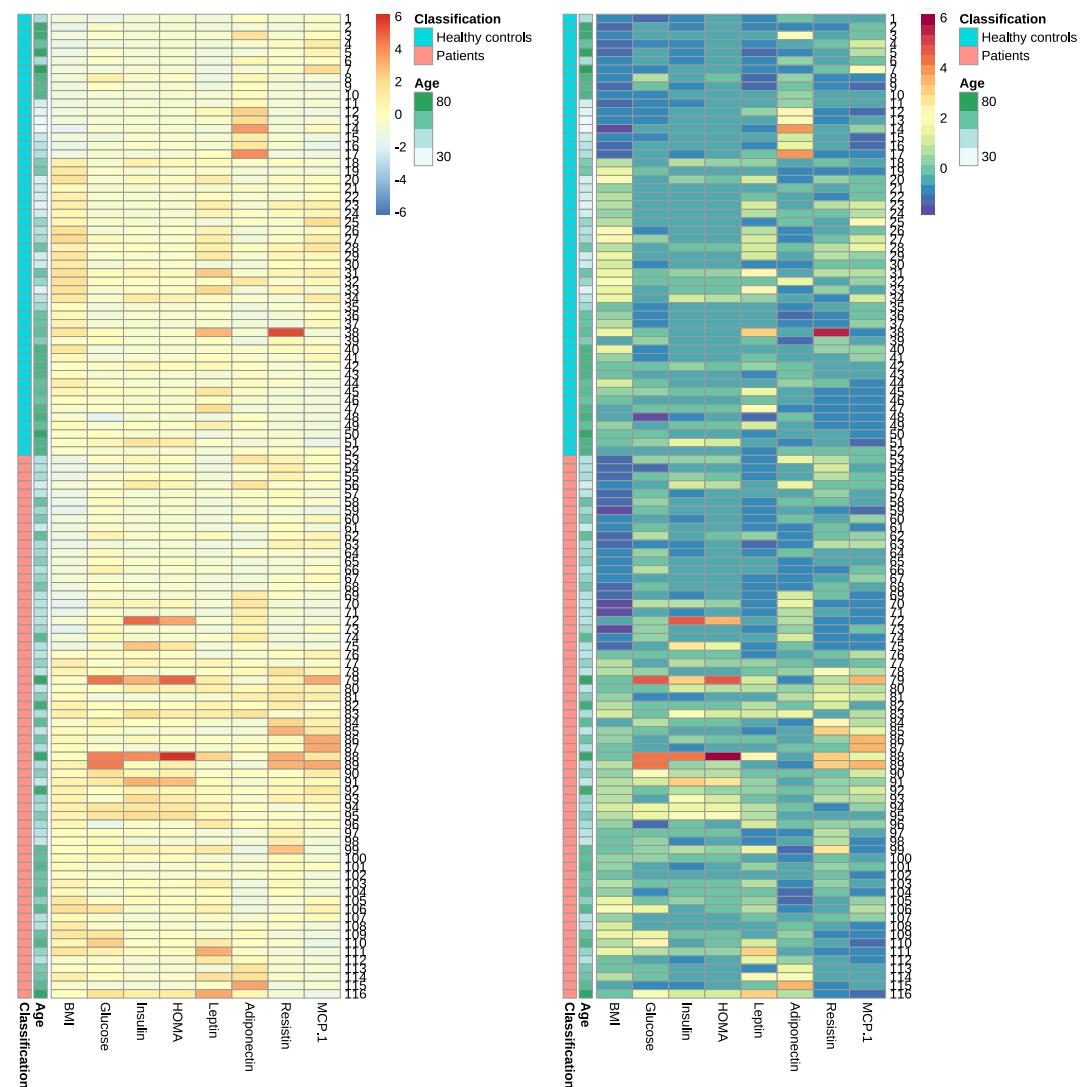
8.32 (群集分析)

資料來源 (UCI): <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>
Breast Cancer Coimbra Data Set, clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.

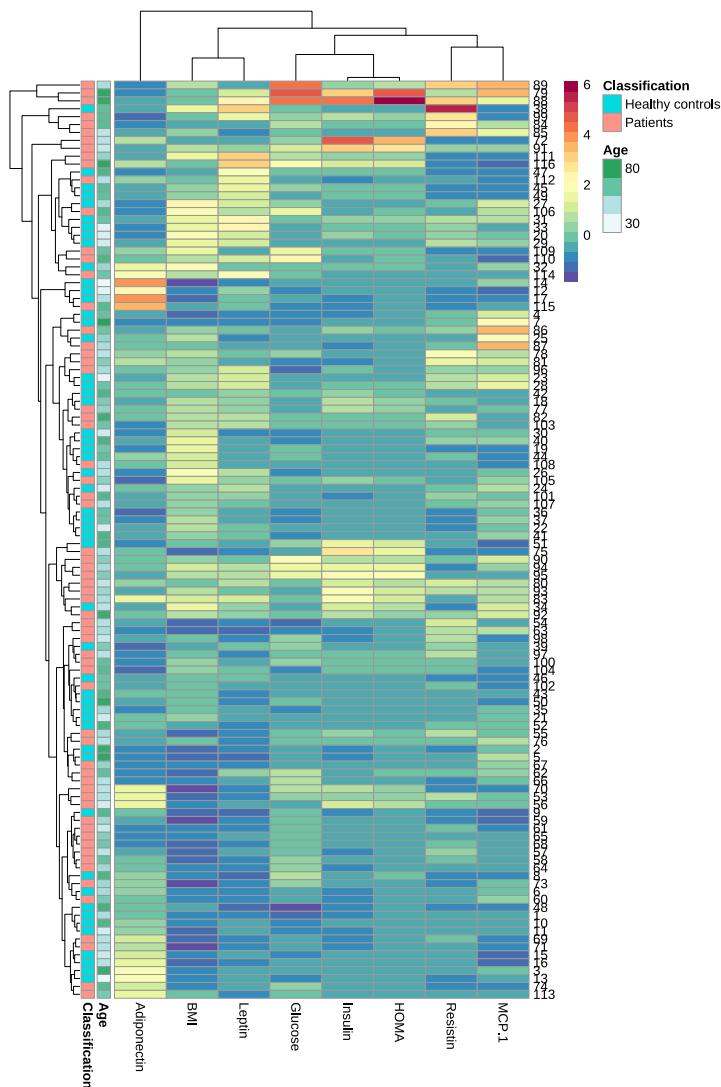
此資料共有 116 人，10 個變數。選取其中 8 個連續變數為解釋變數，記作 $X_{116 \times 8}$ ，(Quantitative Attributes: BMI (kg/m²) · Glucose (mg/dL) · Insulin (μU/mL) · HOMA · Leptin (ng/mL) · Adiponectin (μg/mL) · Resistin (ng/mL) · MCP-1(pg/dL); 剩餘 2 個變數為反應變數，其中 1 個為類別變數 Labels (1=Healthy controls · 2=Patients)，另 1 個為連續變數 Age (years)。

- 以 R 套件 pheatmap 畫出此資料 ($X_{116 \times 8}$)(列及欄皆未排序) 的熱圖，並於此熱圖旁加上 Labels 及 Age 之色條。(請各選合適之色階，需先進行變數標準

化，圖上之列及欄位名稱皆需清晰可辨識)



- (b) 同上小題，對此資料 ($X_{116 \times 8}$) 進行階層式群集分析 (two-ways hierarchical clustering, complete-linkage)，以 pheatmap 畫出相對應之熱圖。(要求：人與人 (rows) 的距離量測尺度為歐式距離，變數間的距離量測尺度為「 $1 - \text{相關係數}$ 」。)



- (c) 同上小題，利用 R 套件 `clValid`，對此資料 (116 人) 進行群集驗証 (cluster validation · internal and stability)，以下列三種分群方法來比較：K-means, PAM, and the hierarchical clustering。

8.33 (群集分析 + 線度縮減)

資料: Mice Protein Expression Data Set from UCI Machine Learning Repository

<http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression>。

變數資訊請參看 Attribute Information。其中 `class` 為類別變數 (共 8 個類別); 欄位 2 到欄位 78(名稱: DYRK1A_N ~ CaNA_N) 為 77 種蛋白質 (proteins) 的表現量。

- 讀入資料 (`Data_Cortex_Nuclear.xls`)，並將有遺失值的變數 (蛋白質) 刪除，不進入分析。
- 以 Fisher's criterion $BW = (BSS(j)/WSS(j))$ 選取前 50 個最能區別 `class` 的蛋白質變數 j 。(記為 `CortexNuclear2`, 其中蛋白質變數欄位已按 BW 排序。)
- 以 `heatmap` 對 `CortexNuclear2` 做群集分析，距離量測指標為 $d_{ij} = (2 -$

$2r_{ij})^{1/2}$, 其中 r_{ij} 為 i th 蛋白質 (老鼠) 和 j th 蛋白質 (老鼠) 的相關係數，階層式群集分析則用 average-linkage。需標上 class 類別變數。表現量色階使用 fields 套件裡的彩虹色。

- (d) 以 LDA 分析 CortexNuclear2, 畫出維度縮減後的 Mouse(前兩維) 散佈圖，並以 class 為顏色，需加上圖例說明 (legend)。
- (e) 以 PCA 分析 CortexNuclear2, 畫出維度縮減後的 Mouse(前兩維) 散佈圖，並以 class 為顏色，需加上圖例說明 (legend)。

8.34 (SVD 影像壓縮)

請對蒙娜麗莎影像 (*MonaLisa_wiki.jpg*) 進行 SVD 影像壓縮，要求如下：

- (a) RGB 彩色影像轉成灰階影像之公式為: $\text{Grey} = 0.2126 \times \text{Red} + 0.7152 \times \text{Green} + 0.0722 \times \text{Blue}$ 。
- (b) 畫出奇異值之陡坡圖。
- (c) 以一頁 4 張圖 (1×4) 做影像結果呈現，最左為原始彩色影像，往右依序為「原始灰階影像」、取前 5 個最大奇異值之「灰階重建影像」、取您覺得最合適之前 k 個最大奇異值之「灰階重建影像」。各個影像需加上合適之主標題。

8.35 (分類法則)

資料集: Microarray gene expression dataset from Khan et al., 2001.

This dataset (subset of 306 genes) can be obtained from R/Bioconductor package "made4": <https://www.bioconductor.org/packages/release/bioc/html/made4.html>. Khan contains gene expression profiles of four types ("EWS", "BL-NHL", "NB" and "RMS") of small round blue cell tumours of childhood (SRBCT) published by Khan et al. (2001). The `khan$train` data.frame is a subset of Khan that contains 306 genes with 64 arrays.

- (a) 以 Fisher criterion (BSS/WSS) 選出前 30 個 BSS/WSS 值最大之 genes。(C03: 分類法則 Classification，第 35/141 頁)。
- (b) (承上小題) 利用 R 套件 caret，使用下列分類方法 ("knn", "rpart", "rf", "adaboost", "svmRadial", "xgbTree") 預測此資料 SRBCT 之四個子型，並印出各分類方法之 10-fold CV error rates。(請自行選用合適之輸入參數)

8.36 (分類法則)

- 資料集名稱: 皮馬印第安人糖尿病資料 (Pima Indians Diabetes Database)
- 資料來源: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- 資料集檔案: `pima-indians-diabetes-database.csv`
- 資料簡介: 此資料集原始出處為國家糖尿病/消化/腎臟疾病研究所 (National Institute of Diabetes and Digestive and Kidney Diseases)。資料集的內容是皮馬印第安 21 歲以上的女性的醫療記錄，以及過去 5 年內是否有糖尿病。

- 分析目的: 依據 8 個診斷變數, 預測皮馬印第安 21 歲以上的女性是否患有糖尿病, 是二元分類問題。
- 診斷變數說明:
 - Pregnancies: 懷孕次數 (Number of times pregnant)
 - Glucose: 口服葡萄糖耐量試驗中 2 小時後的血漿葡萄糖濃度 (Plasma glucose concentration a 2 hours in an oral glucose tolerance test)
 - BloodPressure: 舒張壓 (Diastolic blood pressure (mm Hg))
 - SkinThickness: 三頭肌組織 (皮層) 褶厚度 (Triceps skin fold thickness (mm))
 - Insulin: 餐後 2 小時血清胰島素 (2-Hour serum insulin (mu U/ml))
 - BMI: 體脂數 (Body mass index (weight in kg/(height in m)²)) (體重/身高)²
 - DiabetesPedigreeFunction: 糖尿病譜系功能
 - Age: 年齡 (歲)
 - Outcome: 反應變數, 是否患有糖尿病 (0= 不發病, 1= 發病)

問題:

- (a) 讀取資料, 印出資料之維度、前後各 5 筆紀錄、資料之摘要及結構。
- (b) 以 ggplot2 套件, 畫出 8 個診斷變數之索引圖, 圖上的點以不同顏色及外型表示兩種 Outcome。
- (c) 以 ggplot2 套件, 畫出 8 個診斷變數於兩種 Outcome 類別之 side-by-side 盒型圖。
- (d) 以 ggplot2 套件, 依據兩種 Outcome 類別, 各畫出 8 個診斷變數之直方圖。
- (e) 以 rpart 套件, 建立決策樹模型, 其中訓練資料集佔原資料 80%, 剩餘 20% 為測試資料集。畫出決策樹, 印出混淆矩陣, 並計算五項評估指標: Recall Rate, Precision, Specificity, Accuracy 及 F1-Score。(set.seed(12345))
- (f) 承上小題, 以 ROCR 套件, 繪出 ROC 曲線圖, 並計算 AUC 指標。
- (g) 以 randomForest 套件, 建立隨機森林預測模型 (80%, 20%)(請自行調整合適之模型參數), 印出混淆矩陣, 並計算五項評估指標: Recall Rate, Precision, Specificity, Accuracy 及 F1-Score。
- (h) 承上小題, 以 pROC 套件, 繪出 ROC 曲線圖, 並計算 AUC 指標。

8.37 (分類法則)

資料說明

- 資料集名稱: Wine Quality Dataset (Red Wine)
- 資料來源: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality> 或 <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>
- 資料集檔案: winequality-red.csv

- 參考文獻: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- 資料簡介: 葡萄酒品質資料是用於多類別 (multi-class) 分類問題，亦可用於迴歸問題。此資料的反應變數為酒的 (sensory) 品質等級，共有 11 級，以 0 至 10 的整數表示，數值愈大等級愈高。每一等級的葡萄酒個數不盡相同。葡萄酒的量測變數 (屬性) 共 11 個 (解釋變數是屬於 physicochemical)(見以下說明)。此資料集有兩個子資料集: 白葡萄酒的觀察個為 4898，紅葡萄酒的觀察個為 1599，無遺失值，兩種皆是 Portuguese "Vinho Verde" wine 的變種。(本習題使用紅葡萄酒資料集)。此資料另可應用 Outlier detection 演算法去偵測極差或極優品質的酒，或以特徵選取方法選取能區別葡萄酒品質的重要化學變數。其它細節可參看「資料來源」。
- 分析目的: 以葡萄酒的化學成份 (chemical measures) 指標來預測葡萄酒品質。
- 變數說明:
 - fixed acidity (固定酸度)、volatile acidity (揮發性酸度)、
 - citric acid (檸檬酸)、residual sugar (殘糖)、
 - chlorides (氯化物)、
 - free sulfur dioxide (游離二氧化硫)、
 - total sulfur dioxide (總二氧化硫)、
 - density (密度)、pH (酸鹼度)、sulphates (硫酸鹽)、alcohol (酒精度) 及 quality (score between 0 and 10): 酒的品質。其中 quality 為反應變數，是葡萄酒專家至少 3 次評估的平均值，以 0~10 表示，數值愈大等級愈高。

問題

- (a) 讀取資料 `winequality-red.csv`，命名為 `wine.orig`，印出資料之維度、前後各 5 筆紀錄、資料之摘要及結構。
- (b) 將 `wine.orig` 中的反應變數 `quality` 數值，以 {[0–3]、[4–7]、[8–10]} 之標準，分割成 poor、normal、excellent 三個等級。做為以下接續分類問題的反應變數。請印出分割後的三個等級，各有多少觀察值的酒。
- (c) 以 `ggplot2` 套件，畫出 11 個變數之索引圖，圖上的點以不同顏色及外型表示三種酒品質等級。
- (d) 以 `ggplot2` 套件，畫出 11 個變數於三種酒品等級之 side-by-side 盒型圖。
- (e) 資料轉換:
 - i. 將 `wine.orig` 中的 11 個解釋變數進行標準化後，另儲存成一資料框，命名為 `wine.std`。
 - ii. 將 `wine.orig` 中的 11 個解釋變數進行 [0, 1] 轉換後 (即將每一變數之範圍轉換至 0–1 之間)，另儲存成一資料框，命名為 `wine.unit`。
- (f) 將上述三個資料集，各自分割為訓練集 (佔原資料 80%) 及測試集 (剩餘的

20%)，應用以下列兩種分類模型 (自行選取合適的參數後，固定同一組參數應用於三個資料集)，計算測試集的 10-fold CV 分類錯誤率，以進行比較。

- i. 以 e1071 套件中的 svm 指令，建立支持向量機 (SVM) 分類模型。
- ii. 以 xgboost 套件，建立極限梯度提升 (XGBoost) 分類模型。

8.38 (維度縮減 + 分類法則)

資料來源 (UCI): <https://archive.ics.uci.edu/ml/datasets/Wine>

資料檔: wine_178x13.txt。

Wine Data Set 資料集是在義大利不同地點所生產的三種類別葡萄酒資料 (此類別變數記做 y)，共 178 筆，具有 13 個變數 (特徵)(此 13 個變數記做 X)，皆為量測酒之化學成份所得到的數值。

- (a) 利用 R 套件 pheatmap，畫出此資料的熱圖 (請選用合適之色階)，並以不同顏色標註三種酒之類別。(註: 需做欄位及列之排序 (群集分析))。(參考: <https://davetang.org/muse/2018/05/15/making-a-heatmap-in-r-with-the-pheatmap-package/>)
- (b) 對此資料進行四種維度縮減方法: MDS、PCA、ISOMAP 及 SIR，並畫出維度縮減後的資料於二維平面的投影散佈圖，需用顏色標註三種酒之類別 (註: 各維度縮減方法，請自行選用合適之輸入參數)。(維度縮減講義: 105/144)
- (c) 承 (a) 小題，畫出 co-ranking 矩陣圖，並計算 LCMC($K = 7 \sim 11$)。
- (d) 以下程式碼是利用 R 套件 e1071 中之 SVM 分類器 (使用預設之參數)，將資料 (y, X) 建立一分類模型後，計算分類正確率之範例。(註: 此處不是預測正確率)

```
library(e1071)
attach(iris)
X <- iris[, 1:4]
y <- iris[, 5]
model <- svm(X, y)
pred <- predict(model, X)
accuracy <- sum(diag(table(pred, y)))/length(y)
accuracy
```

以 SVM 計算 Wine Data Set 之分類正確率。

- (e) 承 (b) 小題，以維度縮減後的資料 (依序使用 1 維 ~10 維) 進行 SVM 建模，並計算分類正確率。以 PCA 為例，若維度縮減後的資料變數為 $\{PCA_1, PCA_2, \dots, PCA_{10}\}$ (這裡僅取至第 10 維)，則取 $\{PCA_1\}$ 建模，計算分類正確率；再取 $\{PCA_1, PCA_2\}$ 建模，計算分類正確率，依此類推至 10 維。將結果畫成一個線圖，橫軸為「維度:1, 2, \dots, 10」，縱軸為分類正確率，圖形裡 3 條線 (需為不同型式之線條) 各代表四種維度縮減方法於不同維度的 SVM 分類正確率。(維度縮減講義: 143/144)

8.39 (關聯性分析)

資料說明:

- (a) 資料集名稱: Swiss Fertility and Socioeconomic Indicators (1888) Data
- (b) 資料來源: swiss {datasets}
- (c) 資料集檔案: data(swiss)
- (d) 資料簡介: Switzerland, in 1888, was entering a period known as the demographic transition (人口轉變的時期); i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries. Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. A data frame with 47 observations (地區) on 6 variables, each of which is in percent, i.e., in [0, 100]。
- (e) 變數說明:
 - i. Fertility (Ig): common standardized fertility measure (標準化生育率)
 - ii. Agriculture (%): of males involved in agriculture as occupation (以農業為職業的男性比例)
 - iii. Examination (%): draftees receiving highest mark on army examination (被徵召入伍者於軍隊考試測驗獲得最高分的比例)
 - iv. Education (%): education beyond primary school for draftees (被徵召入伍者有受基礎教育以上的比例)
 - v. Catholic (%): 'catholic' (as opposed to "protestant") (信仰天主教的比例) (相對於“新教”)
 - vi. Infant.Mortality: live births who live less than 1 year (活產嬰兒活不到 1 年的嬰兒死亡率)

問題:

- (a) 將 swiss 資料集之各變數離散化，例如取「低、中、高」之等級 (自行決定切點)，轉成 R 之"transactions" 類別。
- (b) (承上小題) 利用 apriori {arules} 找出依 support 排序之下，前 10 個關聯法則。(請自行選用合適之輸入參數)
- (c) (承上小題) 利用 apriori {arules} 找出依 lift 排序之下，前 10 個關聯法則。(請自行選用合適之輸入參數)
- (d) 各依照上小題，選出有最高 support 及 lift 數值的 frequency itemset 子集合，列出前 10 個與前述 frequency itemset 的 lift 最高的 Frequent Itemsets，並以指令 itemFrequencyPlot 繪出。

8.40 (關聯性分析)

有一檔案 product_by_user.RData 紀錄某商店之產品消費紀錄，包含兩欄位：使用者代碼 (User) 及購買之產品代碼 (Product)。

- (a) 請將 product_by_user.RData 載入 R。所載入之物件是哪一種 R 資料類別？請列印前 5 筆紀錄。

-
- (b) 將產品 (Product) 轉成交易 (transactions) 之資料格式，並列印出前 5 筆紀錄。
 - (c) 使用 `apriori {arules}` 進行關連分析 (自行設定 `support` 或 `confidence` 或 `lift` 之值)。列印出前 5 筆 `support` 最大之關連法則。
 - (d) (承上小題) 列印出前 5 筆 `confidence` 最大之關連法則。
 - (e) (承上小題) 選取出 `lift` 大於 1.5 之關連法則，並依 `support` 值大小排序，列印出前 10 筆關連法則。(若 `lift` 大於 1.5，找不出 10 筆關連法則，則依序降為 1.4、1.3...等等)
 - (f) 利用 `itemFrequencyPlot{arules}`，畫出前 10 個最常出現品項 (Item) 之長條圖。

9 其它

9.1 (爬蟲)

請上 <https://www.amazon.com>，在首頁上方的搜尋列中，鍵入「r statistics」開始搜尋，搜尋結果（預設）以一頁 16 冊書表列於網頁中。請抓取搜尋結果第一頁的資訊，在 R 中以 `data.frame` 儲存。資料需包含「書名，出版年，作者，價錢」。於 R 中，列印抓取之結果。（若講義上之技能，不足以解決這個問題，請盡力想辦法做～）（也可以換成「<http://search.books.com.tw>」，搜尋「R 語言」）

9.2 (R Markdown 文件)

(a) 資料

- i. 資料名稱: 小麥種子 (seeds Data Set)
- ii. 資料來源: <https://archive.ics.uci.edu/ml/datasets/seeds>
- iii. 資料說明: 有三個不同品種的小麥種子: Kama, Rosa 和 Canadian，每一品種小麥皆隨機選擇出 70 顆麥粒，接受檢測。每顆麥粒量測以下 7 種幾何屬性: (1) 面積 (area, A); (2) 周長 (perimeter, P); (3) 緊密度 (compactness $C = 4\pi A/P^2$); (4) 麥粒長度 (length of kernel); (5) 麥粒寬度 (width of kernel); (6) 不對稱係數 (asymmetry coefficient); (7) 麥核溝長度 (length of kernel groove)。

(b) 第一部份

- i. 建立文章主標題 (title) 名為「電概期末考」的 R Markdown 文件，檔名為「學號-姓名-R-FinalExam.Rmd」。輸出格式為「html」。建立子標題名為「第一題」。建立子子標題名為「(a) 建立文件」。資料來源: <<https://archive.ics.uci.edu/ml/datasets/seeds>>
- ii. 建立子子標題名為「(b) 資料說明」。輸入資料說明。
- iii. 建立子子標題名為「(c) 資料摘要」。讀取資料，印出資料摘要。
- iv. 建立子子標題名為「(d) 各小麥品種的麥粒長度直方圖」。以 `ggplot2` 套件，畫出各小麥品種的麥粒長度 (length of kernel) 之直方圖。
- v. 建立子子標題名為「(e) 麥粒長度與麥粒寬度之散佈圖」。以 `ggplot2` 套件，畫出麥粒長度 (length of kernel) 及麥粒寬度 (width of kernel) 之散佈圖。其中圖上的 (三個不同符號) 點以紅、綠、藍三個顏色代表三個不同品種的小麥種子。
- vi. 建立子子標題名為「(f) 各變數盒形圖」。以 `ggplot2` 套件，畫出此資料 7 個變數的 side-by-side 盒形圖。（提示: `seeds.st <- stack(seeds)`）
- vii. 建立子子標題名為「(g) 各變數標準化後之 QQplot」。將每個屬性變數做標準化 $z = (x - \bar{x})/s$ 。畫出每個變數的 QQplot。（Base graphics package）

(c) 第二部份 (建立子標題名為「第二題」。打出下列各數學式)

- i. Poisson 極限定理 (Poisson Limit Theorem):

給定常數 λ ，且 $X_n \sim Binomial(n, \frac{\lambda}{n})$ ，則

$$P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \rightarrow \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{as } n \rightarrow \infty$$

亦即，Poisson 分配是二項式分配 $Binomial(n, \frac{\lambda}{n})$ 的一個極限分配。

ii. 函數

$$f(x) = \begin{cases} |x^2 + x|, & x < 0, \\ \sin(x), & 0 \leq x < 3, \\ 3e^x, & x \geq 3. \end{cases}$$

9.3 (參考講義: T02-hmwu_R-FinTech-1.pptx 及程式碼 T02.R)

用 R 程式自動補捉上証綜指 2012 年日 K 線圖中出現「黃昏之星」形態的日期，並繪製所找出日期附近(前後五天)的 K 線圖。資料集為「SSEC2012.csv」。「黃昏之星」形態的描述如下：

- 對連續三天的日交易資料進行分析。
- 描述蠟燭實體：第一天的收盤價高於開盤價，即描述紅色蠟燭實體，紅色實體要足夠大；第二天的收盤價和開盤價大致相等，兩者差別控制在一個範圍內；第三天，綠色蠟燭實體用收盤價低於開盤價來定義，兩者的差值要大於等於第一天收盤價與開盤價差值的一半。
- 定義十字星實體位置：第二天的收盤價和開盤價均需大於第一天的收盤價和第三天的開盤價。
- 定義上漲趨勢：用收盤價來表示股票的收益率，收益率為正表示上漲。

9.4 (動量交易策略) 資料檔 pufa-2014.csv 為浦發銀行股票 2014 年日度交易資料。

- (a) 請用兩種方法(作差法、作除法)分別計算 6 日動量值、30 日動量值和 90 日動量值。(印出前後各 6 筆紀錄)
- (b) 繪製浦發銀行股票 2014 年 1 月到 3 月份的日 K 線圖，並在 K 線圖中繪製收盤價曲線，在 K 線圖下方繪製 6 日、30 日動量值曲線。
- (c) 用 R 語言撰寫動量指標交易策略，交易策略如下：
 - i. 若當期動量值大於 0，市場的上漲趨勢較大，signal 為 1，(信號出現時為第 1 期)，第 2 期買入股票。
 - ii. 若動量值小於 0，市場的下跌趨勢較明顯，signal 為 -1，賣出股票。

分別用 6 日動量值、30 日動量值和 90 日動量值制定上述交易策略，計算並比較這三個動量買賣預測準確率。

9.5 (RSI 相對強弱指標)

資料檔 jtyh-2013.csv 為中國交通銀行股票 2013 年日度交易資料。

- (a) 繪製中國交通銀行股票日 K 線圖。

- (b) 計算 6 日 RSI、30 日 RSI 的值，並繪製兩者的曲線圖。
- (c) 設定 6 日 RSI 取值大於 90，為超買區，釋放出賣出信號；6 日 RSI 取值小於 10 時，為超賣區，釋放出買入信號。用 R 語言撰寫程式碼計算這種買賣信號預測的準確率。

9.6 (均線系統策略)

資料檔 Dow-2011-2014.csv 為道瓊指數 (股票名稱為 DJIA)2011 年到 2014 年的日度交易資料。

- (a) 計算 30 日 SMA、30 日 WMA 和 30 日 EMA 的值，用 ggplot 繪製這三條均線圖。
- (b) 繪製 2011 年到 2014 年的月 K 線圖，月度收盤價圖和 3 個月 SMA 的值，結合這三種圖形，運用「簡單動平均交易」策略，用 R 語言補捉「道瓊指數的買賣點」。