R 練習題

吳漢銘*
國立臺北大學統計學系

January 26, 2018

Contents

*hmwu@gm.ntpu.edu.tw

1 基礎: 物件、輸入輸出

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

(a) 印出系統現在之年月日及時間。

(b) 列出目前的工作目錄。

(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`」。

1.2 安裝套件

(a) 從台大資工 CRAN 鏡射站安裝兩個套件"`cluster, clValid`"，並載入 R。

(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`。

1.3 列出電腦作業系統 (含位元數) 及 R 版本等等系統資訊。

1.4 某學生分析空氣品質資料 `airquality` 之風速 (`Wind`) 與溫度 (`Temp`) 的關係，他採用迴歸分析及變共數分析，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`)

1.5 (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) 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:
$1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \cdots, -\frac{1}{100}$

1.7 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

1.8 (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。

1.9 `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))
```

1.10 由螢幕輸入 2 個數字 (例如: 26，87)，印出其總和。

1.11「`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))。)

1.12 某班學生有一考試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, 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?

1.13 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)

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(1:100, 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 利用 `for`，試計算 $(1 \times 2 \times \cdots \times 1000000)$ 之結果所需要的電腦系統時間。

2.6 (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。

2.7 有一 50 筆成績資料如下

```
score <- sample(1:100, 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」。

2.8 某班學生期中考微積分及線代的成績資料如下:

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

2.9 有一 50 筆課業成績資料如下

```
score <- sample(1:100, 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。

2.10 小銘老師有某班學生之期中考試及加分考試兩筆資料，

```
midterm <- sample(1:100, 50, replace = TRUE)
extra <- sample(1:100, 50, replace = TRUE)
```

成績比例為期中考佔 40%，加分考佔 60%。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。試寫一函式，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) 最後被當之學生比例。

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

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

2.12 某班某科原始成績如下:`orig.score <- sample(1:100, 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。

2.13 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(1:100, length(student.id), replace = TRUE)
BasicMath <- sample(1:100, length(student.id), replace = TRUE)
Rprogramming <- sample(1:100, length(student.id), replace = TRUE)
English <- sample(1:100, 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。

2.14 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=".")
my.p <- dnorm(seq(-3,3,length=100))
set.seed(123456)
score <-sample(1:100, 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。

2.15 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 %
```

2.16 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

2.17 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?

2.18 (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。

2.19 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 %
```

2.20 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.				

2.21 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)

2.22 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, ⋯)

2.23 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。

2.24 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，銀行行員要如何發給鈔票，使用的張數會最少?

(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。

2.25 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`。

2.26 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
```

2.27 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
```

2.28 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。

2.29 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。

2.30 (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)

2.31 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

2.32 身分證字號驗証規則如下: 字母 (ABCDEFGHJKLMNPQRSTUVXYWZIO) 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
台南市	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

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)
=========================================================
F183741875
A148992712
T189179230
P139392302
H146359668
```

2.33 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!
```

2.34 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:

(a) 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)。

(b) 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。

(c) 重覆上述開獎過程，開獎數字與之前已開出之號碼不重覆。

(d) 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。
```

2.35 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 ⇒ 判別 ⇒ 報出幾 A 幾 B ⇒ 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
============================
```

2.36 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，贏的話就可到一根香腸。其中點數計算規則如下:

(a) 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」。

(b) 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。

(c) 若兩兩相同 (如 3344)，取大的對子，即 44，點數和為 8。

(d) 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」。

(e) 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
============================
3345             9
1662             逼基
...              ...
...              ...
5555             通殺
2654             無面
2266             十八
```

2.37 某公路經過 $A, B, \cdots, G$ 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
```

2.38 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(1:100, 10, replace=T); english <- sample(1:100, 10, replace=T)
algebra <- sample(1:100, 10, replace=T); ScoreData <- cbind(math, english, algebra)
```

2.39 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  ...
...
```

2.40 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) $(3,4),(2,1),(6,6)$.

2.41 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)

2.42 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(1:100, 30, replace=T)
```

2.43 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)。

```
> score[38:43,]
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∼59 分之間的同學。

(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?

2.44 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: xxxxxx。

(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 = 體重/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

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。

2.45 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)
```

2.46 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 $(x_1, x_2, \cdots, x_n)$。
  - `y.train`: 長度為 n，是訓練集資料每個觀察值的類別 $(y_1, y_2, \cdots, y_n)$，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 $(x_1^t, x_2^t, \cdots, x_m^t)$。
- 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 $(d_1, d_2, \cdots, d_n)$。
  - 於上述 n 個距離中，選出距離最近的 $k = 5$ 個觀察值 $(x^{(1)}, x^{(2)}, \cdots, x^{(5)})$。
  - 上述 $k = 5$ 個觀察值，其相對應的類別為 $(y^{(1)}, y^{(2)}, \cdots, y^{(5)})$。
  - 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(...)
```

3 繪圖

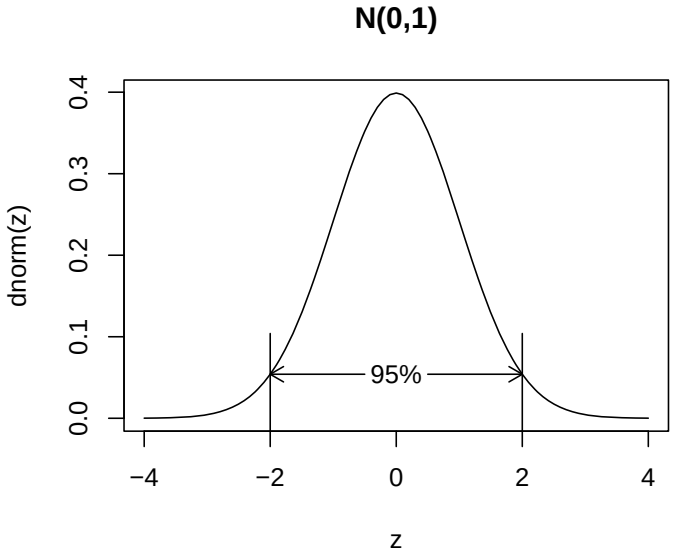
3.1 在標準常態分配的 density function 下，用紅色填滿小於 $z_{0.025}$ 和大於 $z_{0.975}$ 的區域。(提示: `dnorm`, `-1.96`, `1.96`。)

3.2 $z \sim N(0,1)$, 完成下圖。

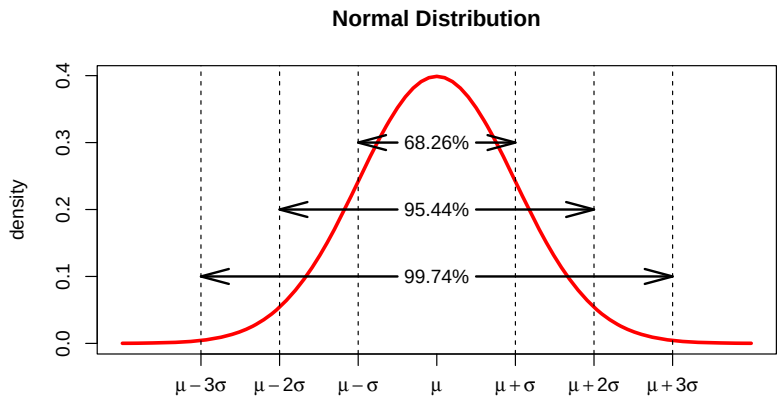
(a) 畫出標準常態分配的 density 圖。

(b) 加上 (紅色) 線段。
(Hint: `segments`。)

(c) 加上 (紅色) 標線及文字。
(Hint: `arrows, text`。)



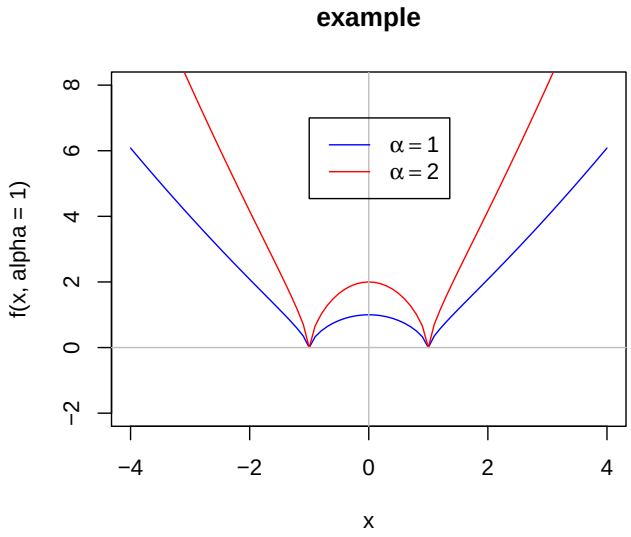
3.3 畫出下圖。(限用兩次 `arrows`，一次 `text`，一次 `abline`，詳見提示。)



提示:

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

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

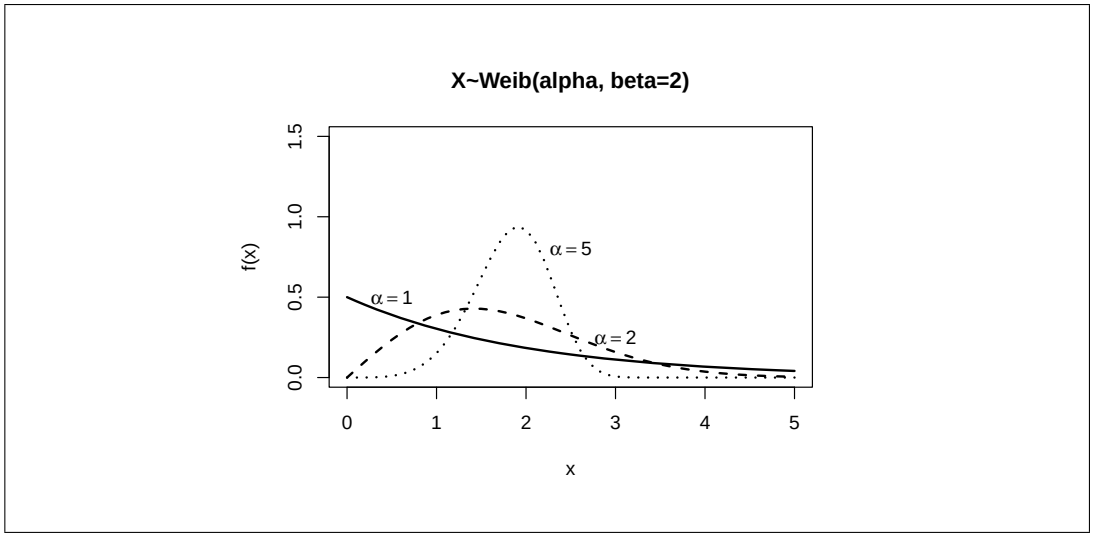


3.5 若隨機變數 X 服從 Weibull 分配 (簡記為 $X \sim Weib(\alpha, \beta)$)，其機率密度函數為

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

(a) 若 `x <- seq(0, 5, 0.1)`，寫一函式計算 $f(x|\alpha = 1, \beta = 2)$ 之值。

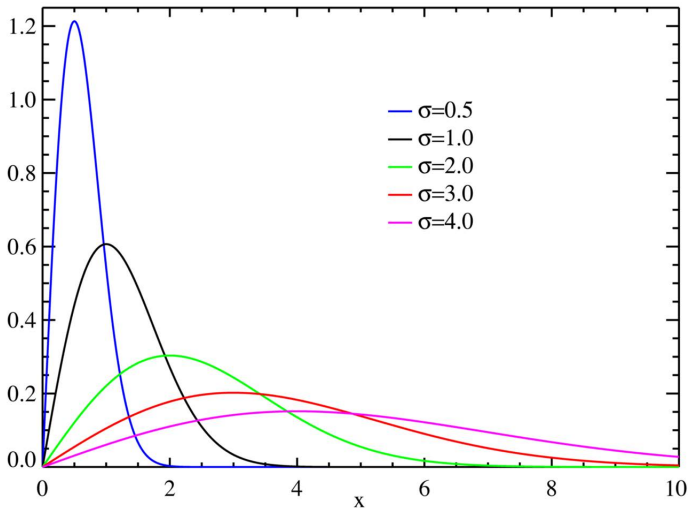
(b) 畫出 $X \sim Weib(\alpha, \beta = 2)$ 之圖形如下:



3.6 The probability density function of the Rayleigh distribution is:

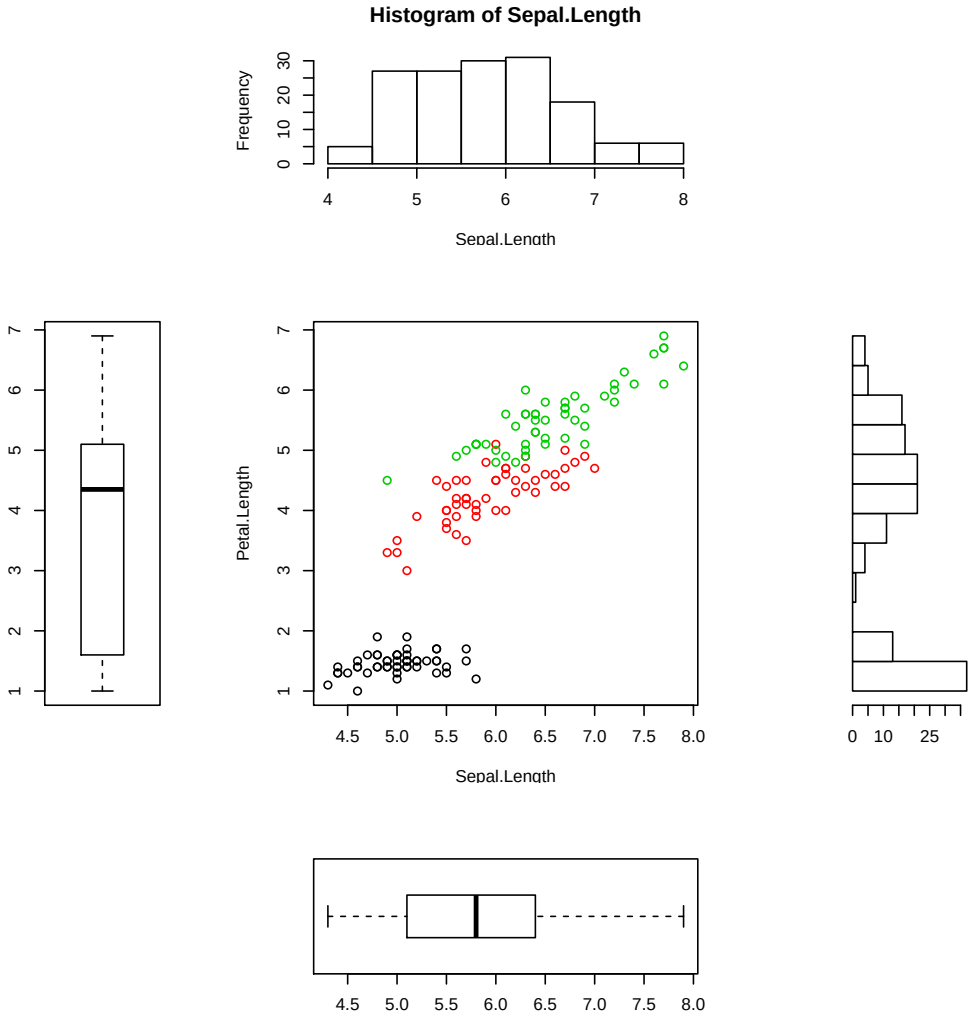
$$f(x;\sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \ x \geq 0,$$

where σ is the scale parameter of the distribution[2]. 畫出此分佈的 Probability density functions 圖形如下:



3.7 利用 `iris` 資料畫出下圖:

[2] `https://en.wikipedia.org/wiki/Rayleigh_distribution`



3.8 (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)。

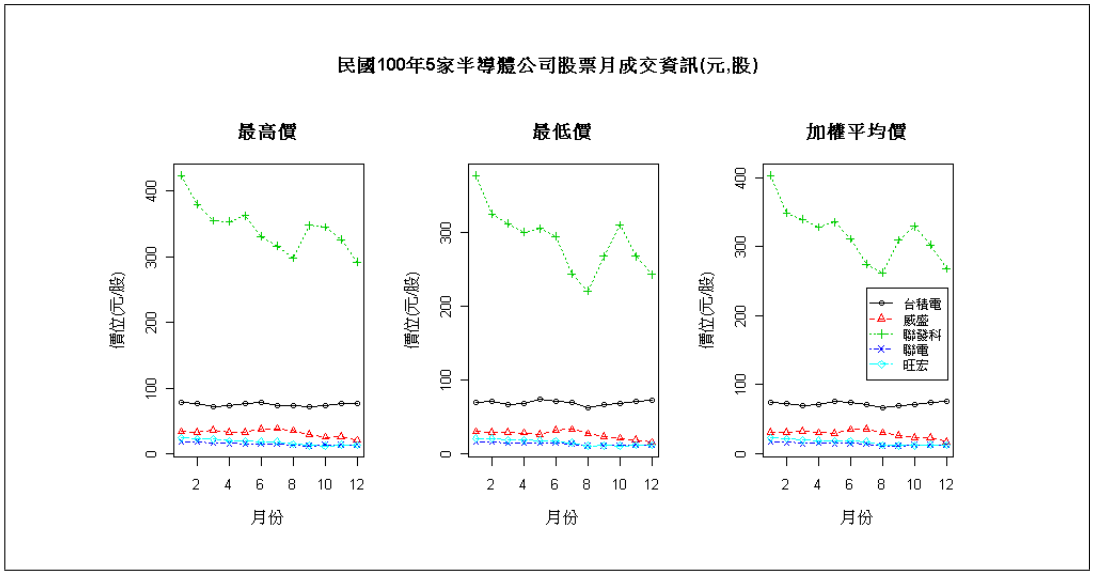
(d) 在這完整的資料中，畫出 `Month=5` 和 `7` 時的 `Wind` (反應變數) 和 `Temp` (解釋變數) 的散佈圖，並加上一條迴歸線 (紅色)。其中 `Month=5` 和 `7` 要有不同的符號 (例如: `a, b`)，不同的顏色 (例如: `blue, green`)。要加 `legend`，要加 `xlab, ylab, main`。

(e) 將最後一張圖存成 `jpg` 檔，然後再 MS Word 裡插入此圖形。

3.9 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，

(a) 請讀入此資料並列印出前 5 筆紀錄。

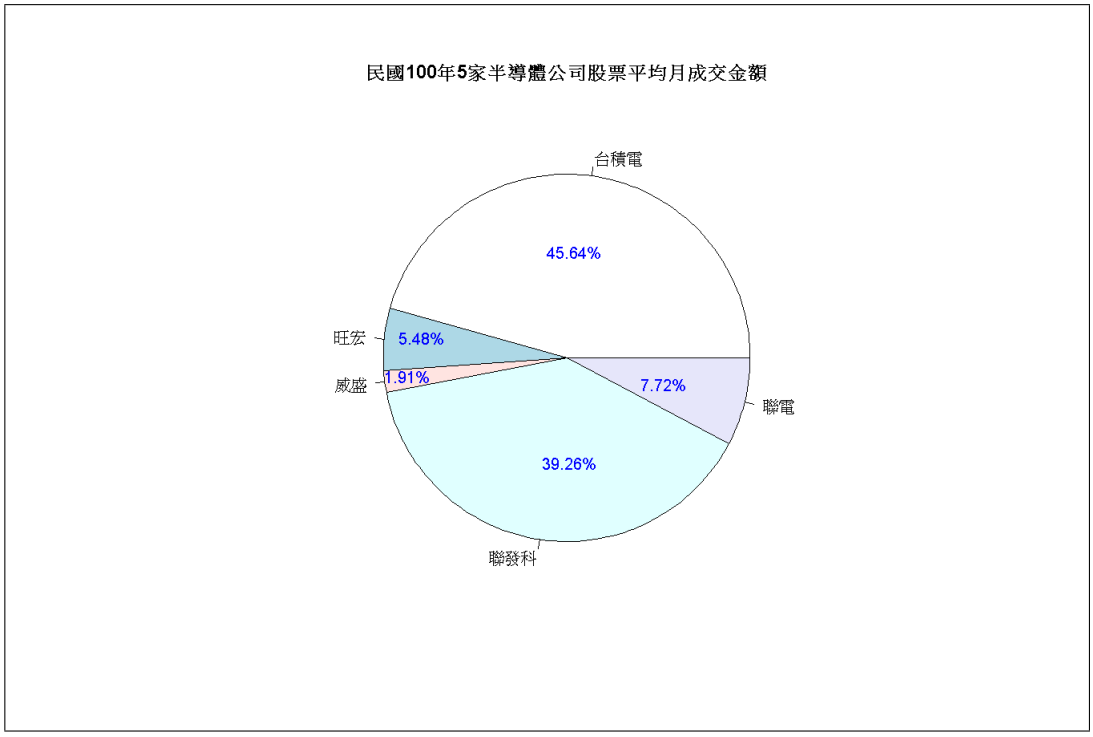
(b) 請繪出以下圖形 (一頁三個圖形)。(提示: `par(..., omi)`, `title(..., outer)`)



(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`)



3.10 教學助理教學評鑑資料 (Teaching Assistant Evaluation Data Set[3])(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

[3]https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation

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 of 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)

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

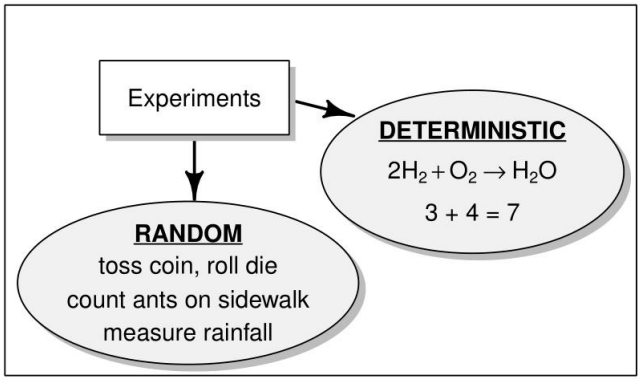
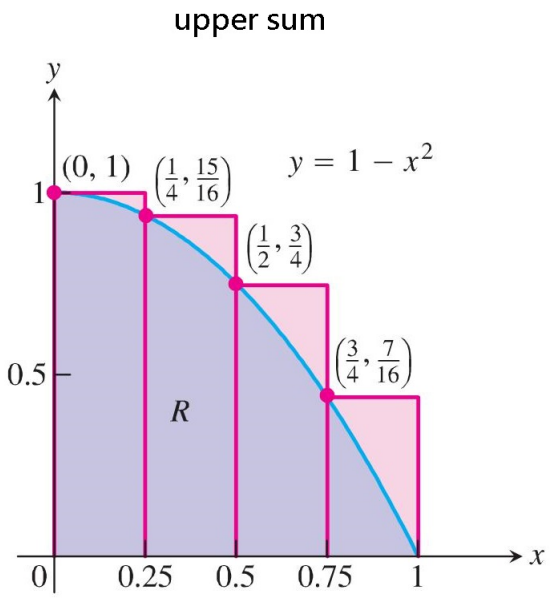
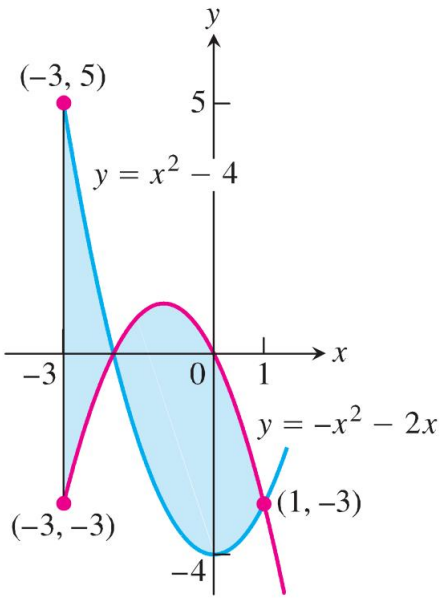


Figure 4.0.1: Two types of experiments

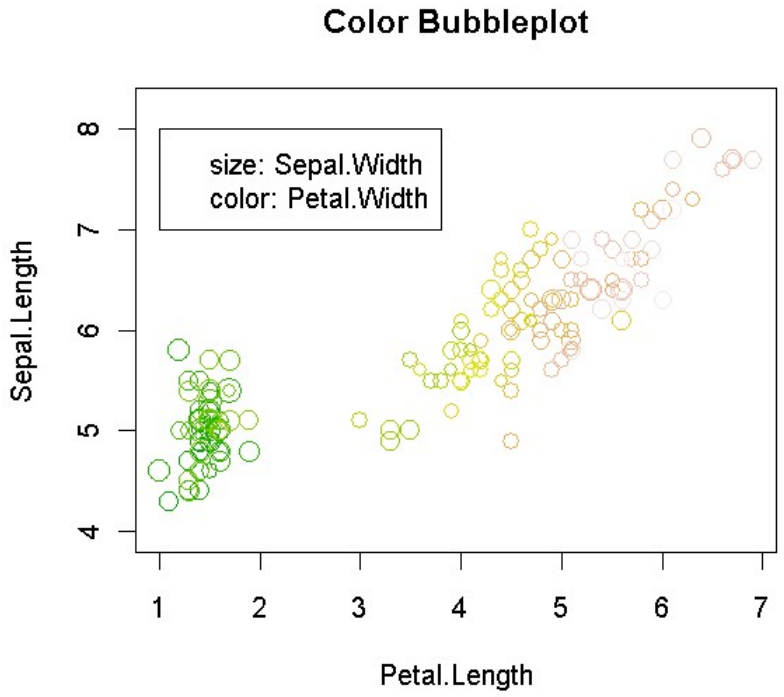
3.12 畫出下圖。



3.13 畫出下圖。(提示: 多邊形 (Polygon), 講義 112/177)



3.14 畫出下圖。



3.15 **(perp 指令練習)** 雙變量 (X_1, X_2) 常態分佈機密度函數定義如下: 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

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

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密度函數之 `R` 函式。輸入為 $(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, cov_{12})$，輸出為 $f(x_1, x_2)$.

3.16 依照下列參數，畫出雙變量常態分佈機密度函數圖。
(可參考: http://tagteam.harvard.edu/hub_feeds/1981/feed_items/177468)

(a) $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, cov_{12} = 0.$

(b) $\mu_1 = \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 9, cov_{12} = 0.$

(c) $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, cov_{12} = 0.99.$

3.17 **(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。

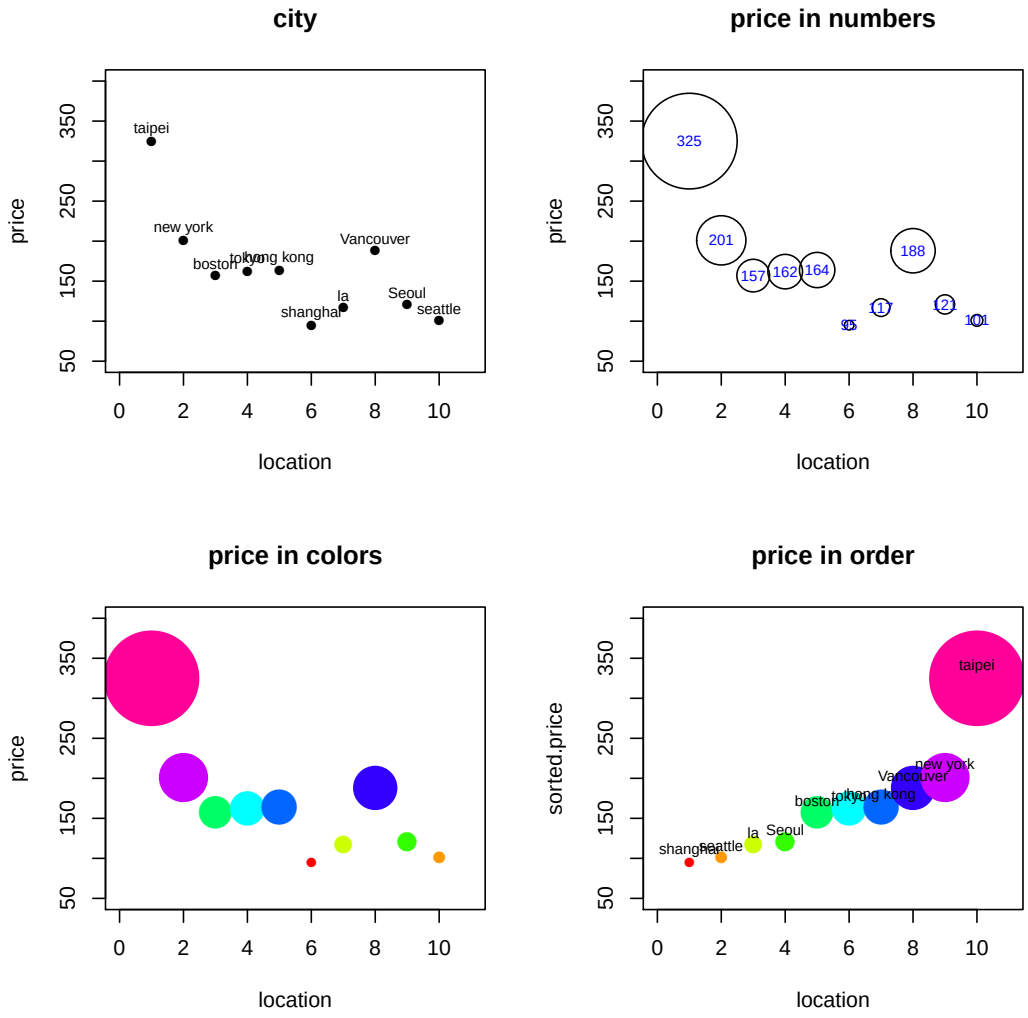
```
> mydata <- data.frame(matrix(sample(0:100, 15*6, replace=T), ncol=6))
> rownames(mydata) <- paste0("student", 1:15)
> mydata
          X1 X2 X3 X4 X5 X6
student1  44 94 63 60 95 89
student2  75 48 23 30 59 37
student3   5 32 93 96 70 74
student4  35 22 49 99 58 19
student5   4  8 46 15 25 23
student6  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"
```

3.18 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。

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

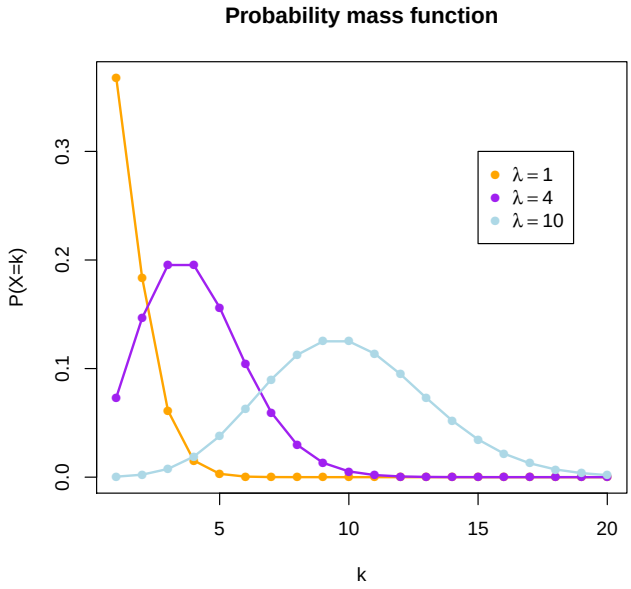
(b) 繪出下列一頁 4 圖 (其中 Bubble plot 之泡泡大小是依據上小題)。
(提示: sort, order, rainbow(10))



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

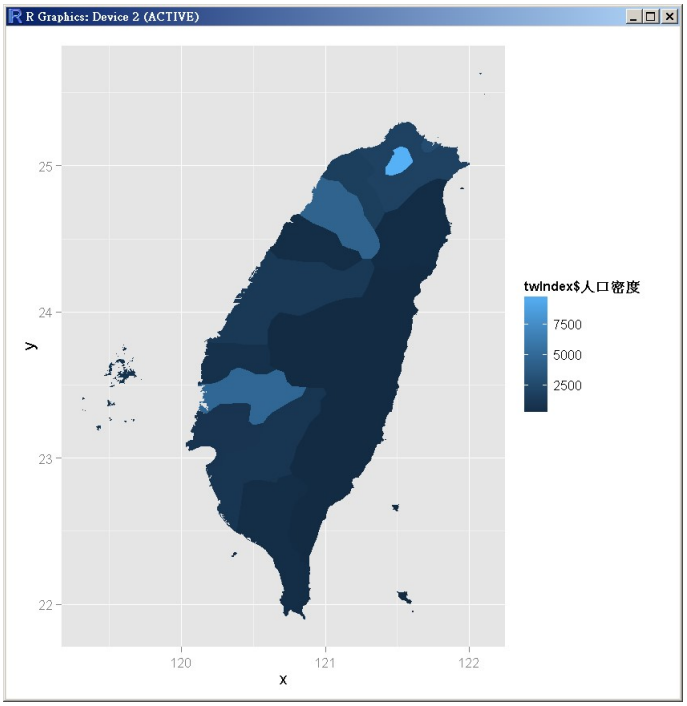
$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

利用 `matplot` 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`)



3.20 **(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`)

3.21 **(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。



3.22 **(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
```

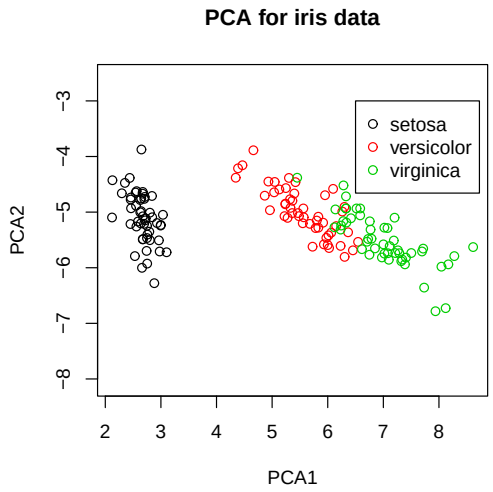
3.23 主成份分析法 (Principal Component Analysis, PCA) for Iris Data

(a) 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)。

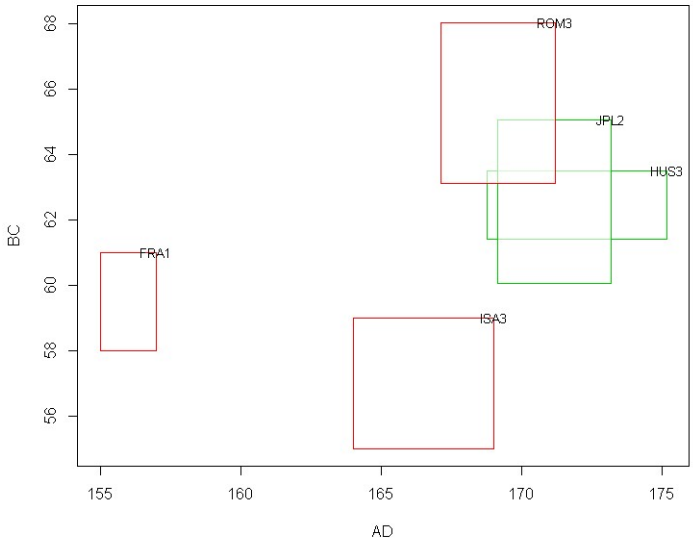
(b) 求出矩陣 `Sx2` 的特徵向量 (eigenvectors) 矩陣 (命名為 `evec` 並印出)。

(c) 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)。

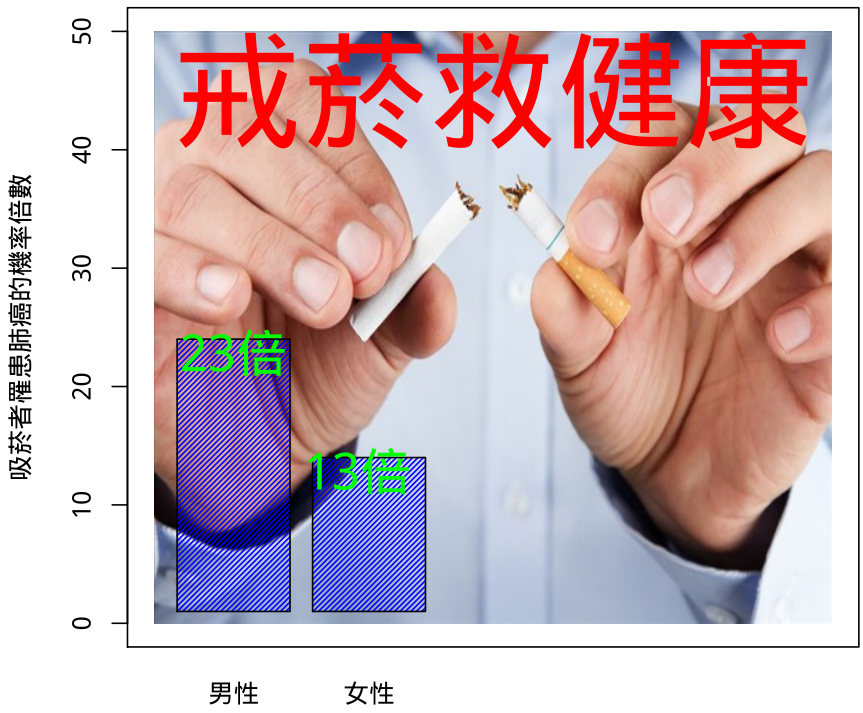
(d) 畫出 x-軸為 `PCA1`, y-軸為 `PCA2` 之散佈圖，並標上 Species 的顏色如下圖。



3.24 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。



3.25 (**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`)



3.26 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)

4 微積分、線性代數

4.1 有三個矩陣如下，計算 (a) AB。(b) $2A + 3C^t$.

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

4.2 寫一函式，使其執行時會要求使用者輸入一 3×3 矩陣，在計算其反矩陣後，回傳原始矩陣及其反矩陣。(提示: `solve`)。請利用以下矩陣做測試:

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

4.3 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 ...
```

4.4 有一數學函數定義在整個實數線上，如下:

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

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   -    -   -    -     -
```

4.5 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，以數值的方法，判別以下給定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 <- ...
....
```

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

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

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
```

4.7 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:

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

- $P = \{x_0 = a, x_1, \cdots, x_{n-1}, x_n = b\}$ 為 $[a, b]$ 之分割 (partition),
- $\Delta x_i = x_i - x_{i-1}$, $i = 1, \cdots, n$,
- $c_i \in [x_{i-1}, x_i]$, $i = 1, \cdots, n$, 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 subinterva1)。

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 $n = 20$ 個子區間,(亦即 $\Delta x_i = (b - a)/n$. 試寫一 R 函式，計算三種黎曼和。(提示: 先產生數列 $\{x_0, x_1, \cdots, x_n\}$，再計算不同取法的 c_i 及 $f(c_i)$)

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

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

my.RiemannSum(0, 2, 20)
$RiemannSum
lower.sum  upper.sum  sum.midpoints
...        ...        ...
```

4.8 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，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。

4.9 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:

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

- $P = \{x_0 = a, x_1, \cdots, x_{n-1}, x_n = b\}$ 為 $[a,b]$ 之分割 (partition),
- $\Delta x_i = x_i - x_{i-1}$, $i = 1, \cdots, n$,
- $c_i \in [x_{i-1}, x_i]$, $i = 1, \cdots, n$, 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 subinterva1)。

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 $n = 20$ 個子區間,(亦即 $\Delta x_i = (b-a)/n$. 試寫一 R 函式，計算三種黎曼和。(提示: 先產生數列 $\{x_0, x_1, \cdots, x_n\}$，再計算不同取法的 c_i 及 $f(c_i)$)

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

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

my.RiemannSum(0, 2, 20)
$RiemannSum
lower.sum  upper.sum  sum.midpoints
...        ...        ...
```

4.10 一個函數 f 從 a 到 b 的定積分定義為: $\int_a^b f(x)\,dx = \lim_{\|P\|\to 0} \sum_{i=1}^{n} f(\bar{x}_i)\Delta x_i$, 其中

- P is a partition of the interval $[a,b]$ with n subintervals:
  $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$.
- Δx_i is the width of for the ith subinterval, $\Delta x_i = x_i - x_{i-1}$.
- $\bar{x}_i$ is a sample point for the ith 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 $\int_{-1}^{3} (2x^2 - 8)\,dx$。令 $f(x) = 2x^2 - 8$，請依底下步驟做答。

(a) 請將區間 $[-1, 3]$ 等分割為 10 等份，列印出子區間的端點值: $x_0, x_1, x_2 \cdots, x_{10}$。

(b) 計算並列印出數列 $\Delta x_i = x_i - x_{i-1}$, $i = 1, \cdots, 10$。

(c) 令 $\bar{x}_i = x_i$，計算並列印出數列 $f(\bar{x}_i)$, $i = 1, \cdots, 10$ 。

(d) 計算 $\sum_{i=1}^{10} f(\bar{x}_i)\Delta x_i$ 的值。

(e) 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。

(f) 呈上題 (e)，求分割數 $n = 50$, $n = 100$, $n = 200$, $n = 2000$, $n = 5000$ 時的答案。

(g) 請利用 `integrate` 指令計算 $\int_{-1}^{3} (2x^2 - 8)\ dx$。

4.11 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 $h = (b-a)/n$，$x_j = a + jh,\ j = 0, 1, \cdots, n,\ a = x_0, b = x_n$. 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。)

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

4.13 假設 $\{x_0, x_1, \cdots, x_n\}$ 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:

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

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

(b) 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    ...      ...
```

4.14 一個連續函數的固定點 (fixed-point) 定義如下:

> Let $g : R \to R$ 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 (`x0`)、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 `x0 = 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 機率與統計

5.1 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。

5.2 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)
...
```

5.3 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，其符合多項式分配。(以下各題假設從多項式分配中隨機抽樣)

(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)

5.4 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。

(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，並在圖上加上一條過原點且斜率為 1 的直線。

(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直方圖。

5.5 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 $\frac{5}{36} = 0.1388889$)

5.6 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。

5.7 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(1:100, 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。

5.8 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) 試寫一函式，計算兩變數之相關係數。

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

5.9 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}), $i = 1, \cdots, n$ 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`)

5.10 截尾平均數 (trimmed mean) 是將一資料排序後，將頭尾拿掉一定百分比 `p` 的觀察值，然後用剩下的觀察值計算平均數。

(a) 試寫一函式，計算截尾平均數。

(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。

5.11 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):

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

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: ...
```

5.12 Compute the values of $f(x) = \dfrac{\sqrt{x^2+100}-10}{x^2}$ 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           ...
======================
```

5.13 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	$B(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	$k=0,1,2,\cdots,n$	np	$np(1-p)$
卜瓦松分佈	$Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$	$k=0,1,2,\cdots$	λ	λ
幾何分布	$G(p)$	$(1-p)^k p$	$k=0,1,2,\cdots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

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:

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

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: $k = 0, 1, 2, \cdots, 100$)

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: ...
```

5.14 計算名目變數 (nominal variable) 的變異分散程度，其中 Index of Qualitative Vari-

ation (IQV) 是一個指標。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)]
```

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

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

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)

(a) 隨機產生 100 個 $Poisson(\lambda)$ 隨機數，畫出其直方圖 (圖標題是 Poisson(λ)，λ 需換成數字)。

(b) 在直方圖上加上 (紅色) 常態分佈曲線 $Normal(\mu = \lambda, \sigma^2 = \lambda)$。

5.16 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)。

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

(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: ($P(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}}) = 1-\alpha$)。(6) p-value: $2 \times P(F \ge f)$。(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/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? ($\alpha = 0.05$)。

5.17 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)。

5.18 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。

5.19 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?

5.20 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?

5.21 (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 $Y = \dfrac{X_1}{X_1 + X_2}$ 是在區間 $[0, 1]$ 之連續均勻分佈。

5.22 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)。

5.23 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 ($\beta_1$) 及截距項 ($\beta_0$) 的估計量如下:

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

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

假設兩變數之資料如下: `x <- iris[,1]`; `y <- iris[,2]`，試計算斜率項及截距項的估計量，並與 `lm(y~x)` 之結果相比較。

5.24 簡單線性迴歸中 ($y = \beta_0 + \beta_1 x + \epsilon$) 之斜率項 ($\beta_1$) 及截距項 ($\beta_0$) 的估計量如下:

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

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

(a) 試寫一函式，計算斜率項及截距項的估計量。

(b) 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。

5.25 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: $g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} - 3$
- Used in SAS and SPSS: $G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$
- Used in MINITAB and BMDP: $b_2 = (g_2+3)(1-\frac{1}{n})^2 - 3$

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。

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

5.26 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 $\{x_1, x_2, \cdots, x_n\}$, $\{y_1, y_2, \cdots, y_m\}$ 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 $P(T > |\mathrm{t}|)$。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
```

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

(a) 令`y <- airquality$Wind`, `x <- airquality$Temp`，以此算出簡單線性迴歸之 MSE。
(Hint: 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: `mean`, $\hat{y}_{j(1)} = \hat{\beta}_{0(1)} + \hat{\beta}_{1(1)}x_j,\ j = 1, \cdots, n$。)

(e) 算出第一點的 Cook's distance ($D_{(1)}$)。
(Hint: $D_{(i)} = \sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2/(p \times MSE)$, 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 ($D_{(i)},\ i = 1, \cdots, n$)) 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`。)

5.28 **Kernel density estimation (KDE, 核密度函數估計)**

Let $x_1, x_2, \cdots, x_n$ 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

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

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。

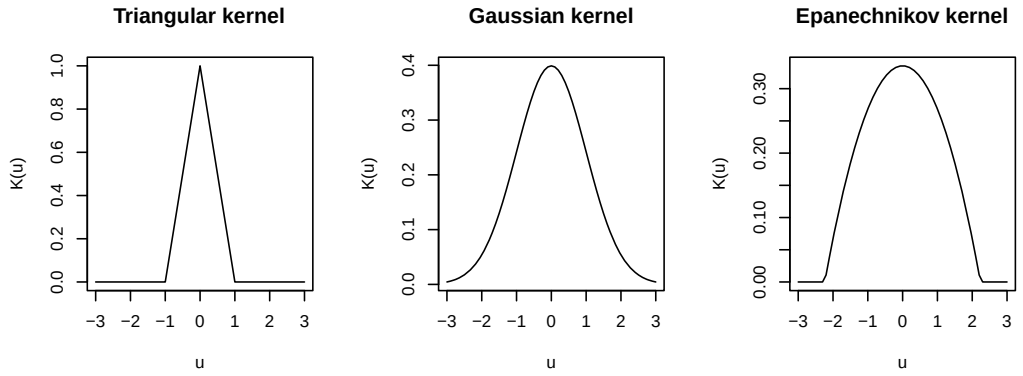
Kernel	Function
Triangular	$K(u) = (1 - \lvert u\rvert)I(\lvert u\rvert \le 1)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$
Epanechnikov	$K(u) = \frac{3}{4\sqrt{5}}(1 - \frac{u^2}{5})I(\lvert u\rvert \le \sqrt{5})$

$$\text{其中} I(|u| \le a) = \begin{cases} 1, & \text{if } |u| \le a \\ 0, & \text{if } |u| > a \end{cases}$$

(a) 完成下列各題。

i. 請將上述三種 kernel function 各寫成一 R 函式。提示: (1) `Triangular <- function(u) { ...`, (2) `if`。

ii. 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 $x_1, x_2, \cdots, x_n$ 為 `xi <- 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
```

5.29 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, $\cdots$, 48 = 女孩; 49, 50, 51, $\cdots$, 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。

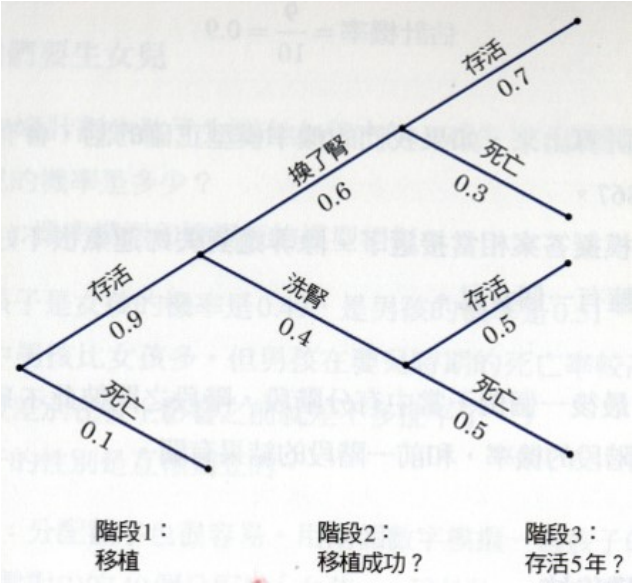
第 4 步：計算機率。若 n 次重複中，有 m 次生女孩。會得到女孩的機率的估計是 m/n。

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	女	男	女	女	男	女	男	男	女	男	男	男	女
	+	+	+	+		+	+		+			+			−	+

5.30 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: 撐過移植手術的占 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; l, 2, 3, 4, 5, 6, 7, 8, 9 = 存活。
  - 階段 2: 0, l, 2, 3, 4, 5 = 移植成功; 6, 7, 8, 9 = 仍需洗腎。
  - 階段 3: 換了腎: 0, l, 2, 3, 4, 5, 6 = 存活五年; 7, 8, 9 = 未能存活五年。
  - 階段 3: 洗腎: 0, l, 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，經 10000 次模擬，計算活過五年機率。

6 資料分析

6.1 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。

6.2 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。

6.3 `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`)

6.4 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`"

(a) 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)

(b) 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)

6.5 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 $\sim$) (也可以換成「http://search.books.com.tw」，搜尋「R 語言」)

6.6 (探索式資料分析) 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:

(a) 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		
	2016	2			
	2016	⋮			
	2016	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.meam[, -(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。

6.7 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 $\mathbf{Z}_{n\times k}$, 其中 n 為酒的個數，$k = 1, 2, \cdots, 10$ 為所取的低維度個數。當 $k = 1, 2, \cdots, 10$ 時，將資料 $\mathbf{Z}_{n\times k}$ 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 $k = 1, 2, \cdots, 10$，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)

$$STRESS = \sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2.$$

6.8 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。

(a) 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。

(b) 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。)

(c) 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 ith 蛋白質 (老鼠) 和 jth 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`)。

6.9 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:
(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}`))