R 基本使用手冊

許湘伶 黃文璋 羅夢娜 編著

國立高雄大學應用數學系

九十三年八月

R 基本使用手冊

作者

許湘伶　國立中山大學應用數學研究所 博士生

黃文璋　國立高雄大學應用數學系 教授

羅夢娜　國立中山大學應用數學系 教授

目　錄

什麼是R　v

第一章　新手上路　1

1.1　軟體取得 1

1.2　安裝程序 3

1.3　開啓方式 5

1.4　操作介面 6

第二章　基本操作指令　9

2.1　輸入與執行 9

2.2　指定資料檔的存放 10

2.3　定義變數 11

2.4　變數的查詢與刪除 12

2.5　說明指令 14

2.6　檔案匯入與匯出 14

2.7　向量與矩陣 16

2.8　使用內建程式庫 18

習題 20

第三章　統計圖　21

3.1　前言 21

3.2 莖葉圖 22
3.3 直方圖 23
3.4 長條圖 23
3.5 盒鬚圖 25
3.6 散佈圖 26
3.7 散佈圖矩陣 28
3.8 常態分位圖 29
3.9 時間序列圖 30
3.10 使用統計軟體R 32
習題 38

第四章 敘述統計量 39

4.1 前言 39
4.2 最大和最小值 39
4.3 中位數、分位數和平均值 40
4.4 變異數和標準差 41
4.5 共變異數 43
4.6 相關係數 43
4.7 使用統計軟體R 45
習題 46

第五章 機率分佈 47

5.1 前言 47
5.2 二項分佈 47
5.3 幾何分佈 48
5.4 波松分佈 49
5.5 Gamma 分佈 50
5.6 卡方分佈 50
5.7 常態分佈 51

5.8 Beta 分佈 52
5.9 t 分佈 53
5.10 $\mathcal{F}$ 分佈 53
5.11 分佈指令表 55
5.12 使用統計軟體R 58
習題 61

第六章 假設檢定 **63**
6.1 前言 63
6.2 單樣本之t檢定 63
6.3 二樣本之t檢定 66
6.4 變異數的檢定 68
6.5 卡方檢定 69
6.6 無母數檢定 70
6.7 使用統計軟體R 73
習題 75

第七章 估計 **77**
7.1 最大概似估計法 77
7.2 區間估計 80
7.3 使用統計軟體R 82
習題 83

第八章 迴歸分析 **85**
8.1 簡單線性迴歸 85
8.2 單因子變異數分析 90
8.3 使用統計軟體R 93
習題 95

第九章　統計品質管制實例分析 **97**
9.1　管制圖 97
9.2　計量管制圖 97
9.2.1　$\overline{X}$管制圖 99
9.2.2　R管制圖 101
9.2.3　S管制圖 103
9.3　計數管制圖 104
9.3.1　p管制圖 105
9.3.2　c管制圖 107
9.3.3　u管制圖 108
9.4　累積和及指數加權移動平均管制圖 111
9.5　使用統計軟體R 113
習題 114

第十章　時間序列實例分析 **115**
10.1 時間序列模型建立步驟 115
10.2 加拿大山貓數量資料分析 116
10.3 太陽黑子數據資料分析 119
10.4 使用統計軟體R 124
習題 126

參考文獻 **127**

附表A **129**

附表B **139**

中文索引 **143**

英文索引 **146**

什麼是R

在R的網頁(The R Project for Statistical Computing, http://www.r-project.org/)上, 可以找到有關R的資料, 僅拮取及整理一些, 以讓大家對R有所了解。

R 是一種用於統計計算和統計圖形的語言及環境。它是一個GNU(為"GNU's Not Unix"的縮寫語, 縮寫恰好又是GNU；它的發音為"guh-NEW") 計畫, 類似於由美國貝爾實驗室的John Chambers等人共同開發的S語言和環境。R可視為是S語言的另一個執行環境, 雖然兩者有一些重要的區別, 但大部分可用S語言執行的指令, 皆可透過R來執行。

R 提供各式各樣的統計(包含線性和非線性模型, 一般的統計檢定, 時間序列分析, 分類以及叢聚等) 和圖解技術, 且具有高度的擴充性。S語言通常被選為研究統計方法的工具, 而R針對此部分提供了一個開放的原始碼。

R的長處之一在於充分設計好且已公開發表的功能, 並包括數學上所需的符號和公式。雖然有一些功能已被一些軟體所取代, 但是使用者可保留完整程式的控制權。

R是一可免費取得的軟體。它提供了適用於Unix、Linix、MacOS和Windows等各種作業系統的版本, 這些版本皆可以免費下載和使用。像S語言一樣, R是一種電腦程式語言, 且允許使用者自行定義新的功能。R的系統大部分是用與S相似的語言所寫成的, 如此可使得R與其他程式語言, 如C, C++, Fortran等, 彼此可在執行程式時作一連結和呼叫的動作。高階的使用者更可直接以編寫C語言的方式來運用R的物件。

與其說R是一個統計系統, 不如說R提供了一個執行統計分析的環境。因為R不僅提供若干統計計算程式, 也可透過程式庫(packages)擴充更多的統計模組。在R的安裝中, 已包含八個基本的程式庫可使用, 另外可透過CRAN(http://cran.r-project,org)獲得其他的程式庫。

R的優點：

- R是一免費的統計計算軟體, 它提供一開放的原始碼, 可在Unix、Linix 、MacOS 和Windows等作業系統上執行。
- R有完整的說明(help)系統, 包括關鍵字查詢、指令查詢以及線上說明系統。
- R的圖解能力很好。
- 學會操作R的使用者, 可輕易地學會商用軟體S-plus, 僅需要了解兩者的差異處即可, 不須重新學習語法。
- 使用者可自行編寫函數, 除了可擴充功能外, 也能滿足不同使用者的要求。一般商用軟體需經過廠商開發, 才能擁有新的功能。
- 已有一些WEB介面發展完成, 如R-php(http://dssm.unipa.it/R-php/), 使用者不須將此軟體安裝在電腦裡, 透過網路亦可使用R。也可上傳自己的檔案進行分析。

R的缺點：

- R是一個有限制的圖形操作介面, 因此新手對R的操作，可能要花段時間才能學會。
- R並沒有商業性的支援, 僅能透過對此軟體愛好的使用者, 針對分析上所需的功能進行開發, 並發表在網站上。
- R的共同語言是一種電腦程式語言, 所以使用者必須學會相關語法才能操作及分析。
- R雖有WEB介面, 但大部分的WEB介面使用者, 仍須會R的相關語法。由於WEB介面的分析得經由網路傳輸, 所以需要一些等待時間才能獲得結果。

第一章

新手上路

1.1 軟體取得

軟體R是一套程式指令, 與S-PLUS相似, 均為統計計算軟體。與S-PLUS最大不同處, 且亦為其優點, 即R為一免費的軟體。任何人只要能透過網際網路連結上R的專屬網站即可取得此套軟體, 並且在自己的電腦上操作。

透過R的專屬網站(www.r-project.org), 即可下載最新的版本(目前版本為rw1090 約20MB), 並可在網站上獲得許多相關資源, 如物件等。

底下我們將詳細介紹如何透過網路取得最新版的統計軟體R。首先連結上專屬網站, 畫面如圖1.1。

接著在左手邊按鈕列中找到Download下為CRAN之按鈕(參考圖1.1), 按下即進到CRAN Mirrors的網頁, 找到適合自己下載檔案的伺服器位置(我們以Taiwan下的http://cran.csie.ntu.edu.tw/ 網頁為例)。之後根據電腦系統配備找到適合自己電腦的連結位置, 進入網頁將所須之軟體下載至個人電腦之中。點選下載鈕後, 可看到如圖1.2之畫面。點選"儲存"可將此套軟體存到個人電腦之中。

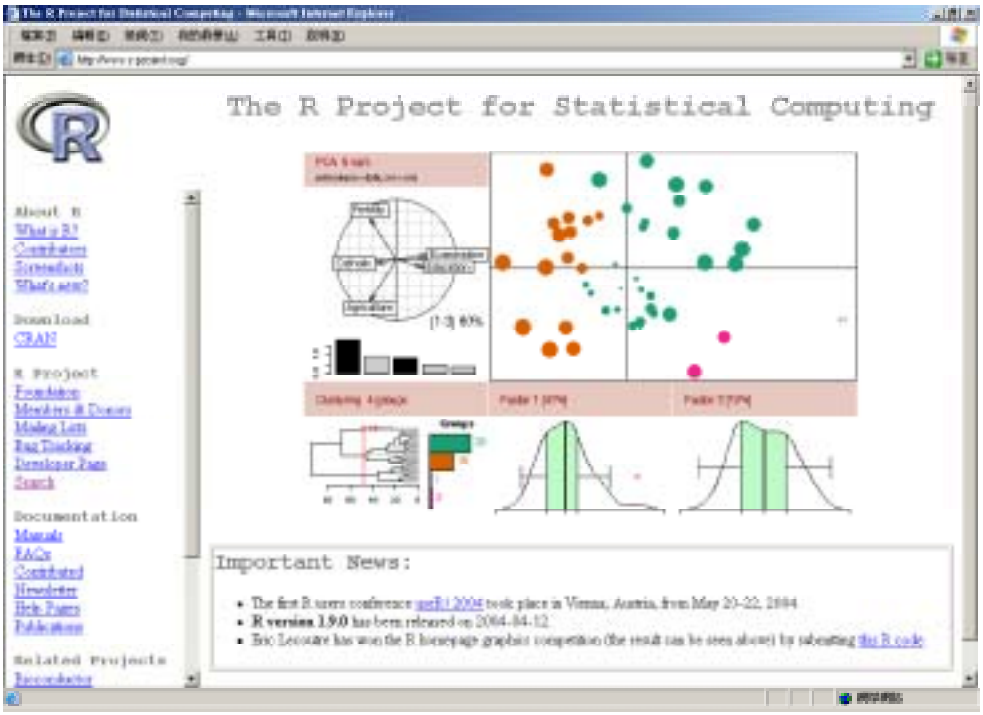


圖1.1 統計軟體R的網站首頁

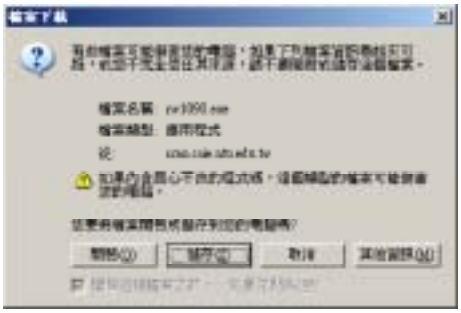


圖1.2 下載統計軟體R的視窗畫面

1.2 安裝程序

下載完成後即可將R正式安裝至電腦中, 我們將按照下述步驟, 完成全部的安裝程序。

首先執行rw1090.exe(即滑鼠左鍵點二次所下載完成的軟體)
, 之後可開啓如圖1.3之歡迎畫面。若欲進入下一步驟之安裝請按“Next”(下一步), 若不想安裝請按“Cancel”(取消) 。

進入下一步驟的安裝程序後需點選“I accept the agreement”(我接受此協議內容), 才可點選“下一步”。接下來的步驟只需按照預設值安裝即可(如圖1.5-1.9)。

最後安裝步驟都完成後會出現如圖1.10的畫面, 點選“finish”(完成) 即可完成所有安裝程序。安裝完成後我們就可以進入下一節來實際操作此套軟體。

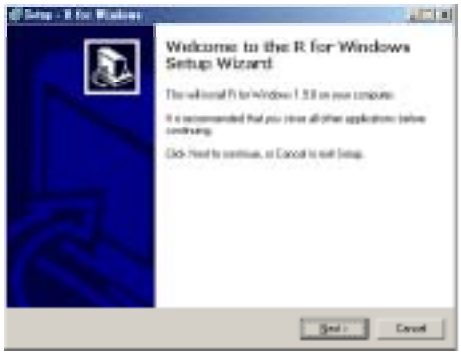


圖1.3 安裝R的歡迎畫面

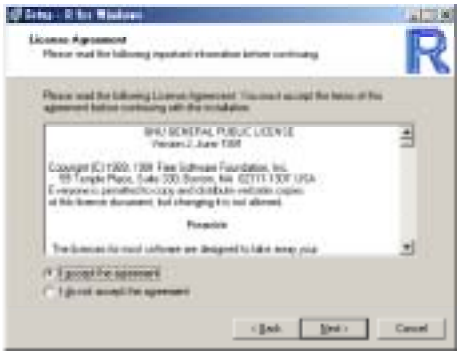


圖1.4 授權協議的畫面

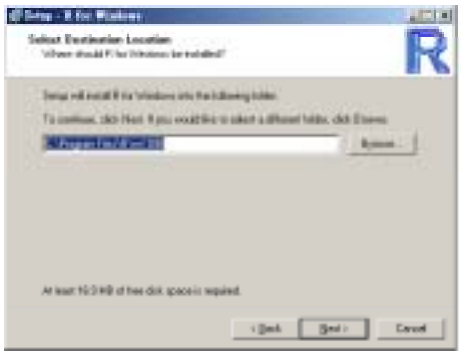


圖1.5 軟體欲安裝位置的畫面

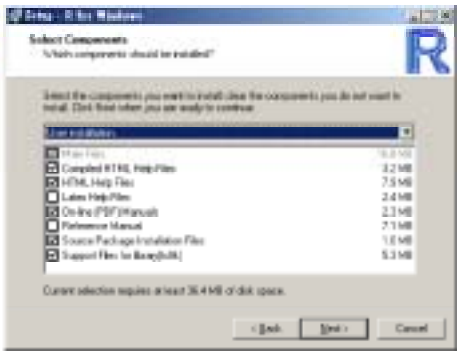


圖1.6 安裝元件選擇的畫面

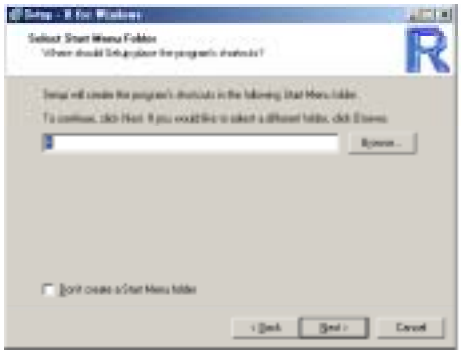


圖1.7 是否在程式集裡放置資料夾

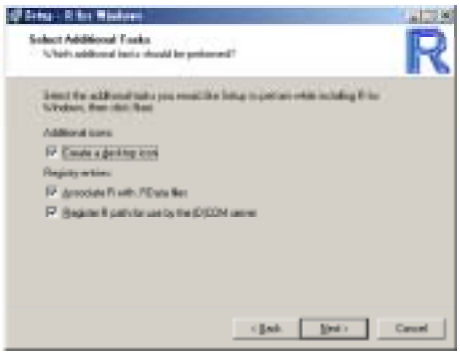


圖1.8 是否安裝額外的工具

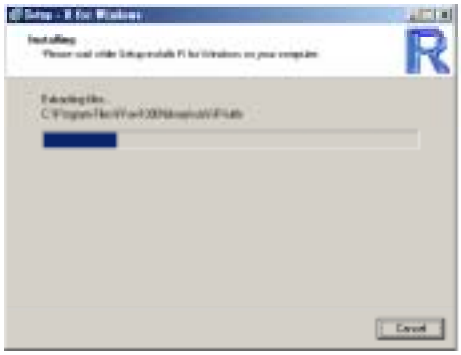


圖1.9 安裝進行中

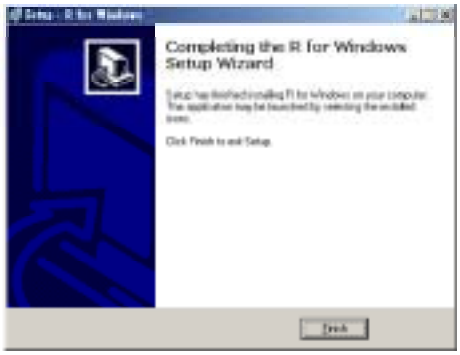


圖1.10 完成所有安裝的畫面

1.3 開啓方式

完成安裝程序後即可開始使用此套軟體, 底下將介紹開啓的方式。

如果在安裝時, 一切步驟均按照內定值來做設定, 那麼會在電腦桌面上出現一個捷徑的圖示(如 R 1.9.0), 利用滑鼠左鍵點二下即可進入起始畫面(如圖1.11)。

若桌面未設有捷徑, 則可由開始進入程式集後, 找到R資料夾的位置, 開啓資料夾內容即可看到執行程式R 1.9.0, 點選後亦可看到如圖1.11之操作畫面。

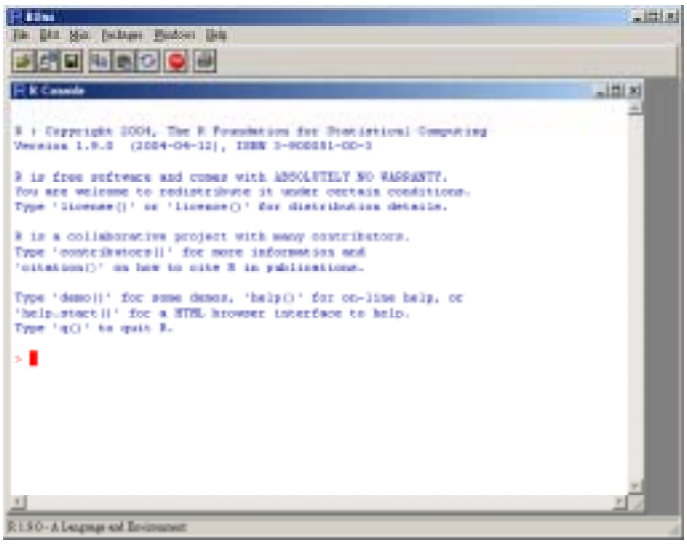


圖1.11 R的操作介面

1.4 操作介面

進入操作畫面之後, 下一步驟就要了解基本的操作方式, 像是檔案的開啓, 指令的搜尋等都要先有初步的概念, 有這些基本的了解對操作上會有不錯的幫助。接下來將針對操作畫面作一簡單的介紹, 整個操作畫面可區分為三大部分, 分別為按鈕列, 工具列和輸入視窗(R console)。

按鈕列共包含六個項目, 分別為檔案, 編輯, Misc, 物件, 視窗及說明, 如圖1.12。

圖1.12 R操作畫面的按鈕列

底下我們將一一介紹其內容。

1. 檔案(file)：包含有匯入和儲存檔案, 匯入和儲存歷史檔案, 更改存取資料夾位置(change dir), 列印(print)和離開此視窗(exit)。

2. 編輯(edit)：包含有複製(copy), 貼上(paste), 清除之前輸入的指令(clear console), 資料編輯(data editor)以及系統設定(Gui preferences)。

3. Misc：停止運算(stop current computation), 列出物件(list objects), 移除所有物件(remove all objects), 列出搜尋路徑(list search path)。

4. 資料庫(packages)：下載程式庫(load packages), 從CRAN下載或更新資料庫等。

5. 視窗(windows)：視窗的排列方式, 分別為階梯狀排列(cascade)和塊狀排列(tile)。

6. 說明(help)：資訊(console), 查詢R的函數(R functions), html畫面說明(html help), 查詢(search help)等。

工具列的部分共包含八個按鈕, 如圖1.13所示。針對每一按鈕一一做說明：

圖1.13 R操作畫面的工具列

1. ：開啓檔案。
2. ：開啓舊檔。
3. ：儲存檔案。
4. ：複製。
5. ：貼上。
6. ：複製和貼上。
7. ：停止運算。
8. ：列印。

在操作畫面上的輸入視窗, 如圖1.14, 一開始介紹了此套軟體的版本, 簡單說明此為免費軟體, 和一些基本需要了解的指令如help。

圖1.14 R操作畫面的輸入視窗

完成所有安裝步驟, 進入操作的畫面, 且有一些基本的了解後, 我們就可以到指令輸入的階段, 下一節將會簡單介紹一些基本常用指令和繪圖等的輸入方法。

第二章

基本操作指令

2.1 輸入與執行

當R處於可輸入指令狀態時, 則會顯示大於的符號, 即“>”。若使用者要執行所輸入之指令只要按“輸入”(enter)鍵, R就會執行所得到的指令。底下給一簡單的例子說明。

$> 1 + 2 \quad \hookleftarrow$

[1] 3

我們以“↵”代表按輸入鍵, 而“[1] 3”為R輸出結果的表示方式。

當指令未完整輸入, 並且按下了輸入鍵, 那麼畫面上大於(>)的符號會轉變成加號, 即“+”, 因此出現了加號的指示就表示指令並未完整輸入, 需將指令輸入完成並按輸入鍵即可執行運算。若要跳出未完成的指令輸入, 則按“Esc”鍵即可輸入新的指令。

2.2 指定資料檔的存放

R有一預設位置來放置儲存的檔案, 預設位置路徑為

C:\Program Files\R\rw1090。

若未做任何更動, 那麼所有的存取皆在此預設位置完成儲存或開啓的動作。

使用者若不想每次都到此預設位置內, 希望能放在自己方便的資料夾中, 可以透過變更檔案的存放位置, 來達到此目的。如何變更位置呢？其方式如下：開啓R軟體之後, 到按鈕列找到“檔案”(file)並按下, 可看到“變更路徑 ”(change dir, 參考圖2.1)執行後, 可得到如圖2.2的畫面。接下來可透過“瀏覽”(browse)按鈕找到想要放置的資料夾。或者使用者可以在輸入列自行輸入資料夾的路徑亦可。

每次當使用者操作完畢之後, 希望關閉視窗, 則會出現如圖2.3的詢問畫面, 問使用者是否要儲存該次操作的內容, 包括輸入的指令和變數。若選擇“是”, 則會將該次資料儲存在設定的資料夾之中, 並存成名為“R.RData”(如 R.RData)檔案。使用者在下次使用時, 只需點選R.RData , 即可讀取前一次儲存的內容。

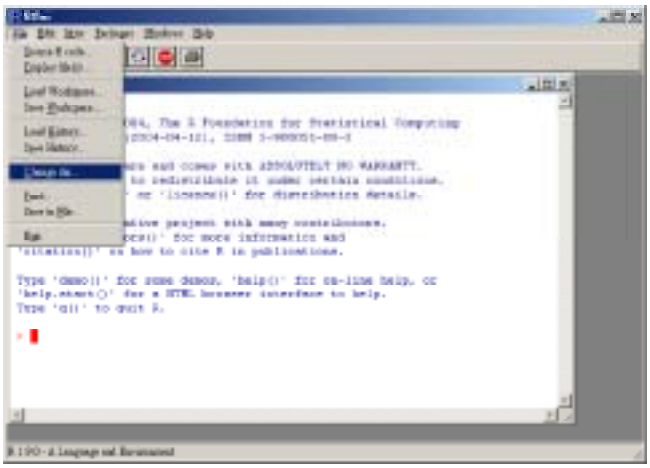


圖 2.1 變更路徑的選項

圖 2.2 變更存放路徑的詢問畫面

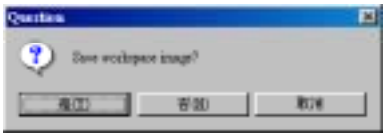


圖 2.3 是否儲存檔案的詢問畫面

2.3 定義變數

一般在做資料分析時, 會收集到一大堆數據, 分析者總是希望數據能精簡, 不必有那麼龐大的數字就能獲得所需要的結果。同樣的, 在編寫程式指令作運算, 會希望是一簡單的代號, 如英文字母或數據名稱縮寫, 來取代一堆數字的輸入。所以如何定義變數就很重要了。

R, 似其他電腦程式語言, 可用“變數”(variable)符號來取代一大堆資料, 讓使用者操作上更簡便。先舉一簡單的例子來說明如何定義變數。例如, 想計算$1+1$。可利用最簡單的方式

$> 1+1$ ↵

[1] 2

除此之外, 可透過變數得到同樣結果。

$> x<-1$ ↵

$> x+x$ ↵

[1] 2

其中符號“<–”為一指定運算符號, 由小於“<”和減號“–”所組成的一個表

示式。表示將數字1“指定給”(assign to) x。若變數僅代表單一數字, 非一整串數字時, 可用“等於”(=)取代“<–”。

雖然透過指定給變數x的方式在此例子上看來較複雜, 但當x所代表的不是單一數字, 而是多筆數據資料時, 經由變數來取代的方式, 就會有很大的幫助。

R在定義一變數並沒有太多的限制條件, 變數名稱除了第一個字必須為英文字母外, 其他位置的字可用“英文字母”(letter, 如a, b等), “數字”(digit, 如1, 2等) 和“點”(dot, 即“.”)。在定義變數有一點需特別注意的是R會將英文字母的大小寫視為不同的變數名稱。

2.4 變數的查詢與刪除

所有使用者所定義的變數, R會將其儲存在一個共同的位置。想要知道使用者已經定義了哪些變數名稱, 或者這些變數包含了什麼資料內容。那麼我們就必須知道如何來查詢變數。查詢變數指令有兩種, 一種為“ls()”, 另一種為“ls.str()”。

首先, 先介紹何為“ls()”。輸入“ls()”指令, 可得如圖2.4。此項指令亦可用“objects()”, 或由之前介紹過的按鈕列上“Misc”的選項“列出變數”(如圖2.5), 此兩種方式皆可達到“ls()”的功能。“ls()”僅將已定義的變數名稱列出, 但是變數包含了什麼資料, 共有幾筆數據等內容皆無法透過此指令得到。因此我們需要另一個指令, 即“ls.str()”, 來獲得更詳細的內容。

當使用者輸入“ls.str()”指令, 可得如圖2.6的畫面。透過這個指令可以看到變數是何種資料型態, 如資料表(data frame, 包含多個變數), 向量, 或數字等, 共有幾筆數據, 並且會列出前幾筆數據資料。

知道如何查詢變數, 接下來另一項指令就是如何刪除使用者不需要的變數。刪除變數指令有“rm(變數名稱)”和“rm(list=ls())”。“rm(變數名稱)”表示刪除使用者指定的變數, 可同時刪除多個變數, 中間以“,”為區

隔符號。“rm(list=ls())”則表示刪除清單(list)為“ls()”所查詢到的所有變數。此兩項指令可參考圖2.7。

圖 2.4 查詢所有變數名稱

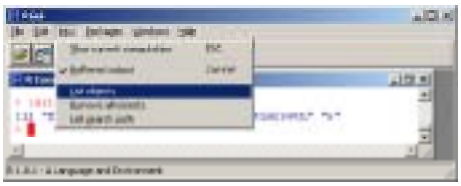


圖 2.5 利用“Misc”查詢所有變數名稱

圖 2.6 查詢所有變數及其相關資料

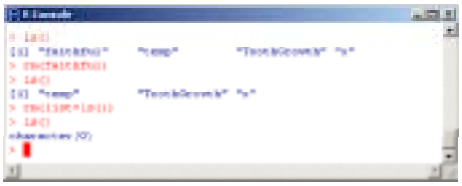


圖 2.7 刪除變數指令

2.5 說明指令

在使用一套軟體而且手邊沒有任何可供參考的書籍時, 常會借助“說明”(help)。R提供了二種主要型式的說明, 一種為完整指令查詢, 一種為關鍵字的查詢。

使用者若已經知道指令(如rm), 但對於指令的細部內容或指令的輸入方式有疑惑時, 可利用“help("指令名稱")”。輸入“help("rm")”指令就會出現關於rm指令的說明及操作的詳細資料, 參考圖2.8。

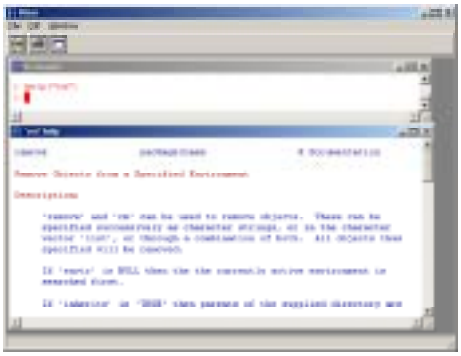


圖 2.8 查詢rm指令

關鍵字查詢分為二種方式, 一種為輸入“help.search("指令名稱或關鍵字")”, 一種為線上查詢(online help)。透過“help.search("指令名稱或關鍵字")”查詢, R會列出指令名稱或指令說明中有包含此關鍵字的所有指令, 如圖2.9。另外一種關鍵字查詢為線上查詢即透過瀏覽器(如Internet Explorer)介面查詢所需要的指令。開啓線上查詢可透過指令“help.start()”, 或由按鈕列上的“說明(help)”的選項“Html help”(參考圖2.10), 可開啓如圖2.11的查詢畫面。

2.6 檔案匯入與匯出

當使用者定義了一些變數並完成資料分析後, 若需要將檔案存成別的格式以利其他軟體(如Excel)的操作。此時就需要知道如何將變數所包含的

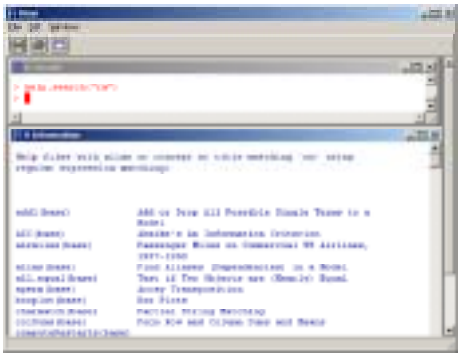


圖 2.9 查詢含有rm所有的指令

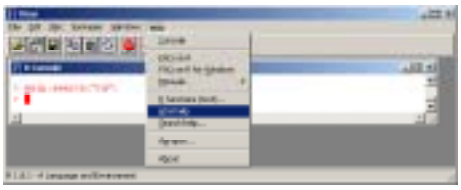


圖 2.10 查詢指令按鈕選項

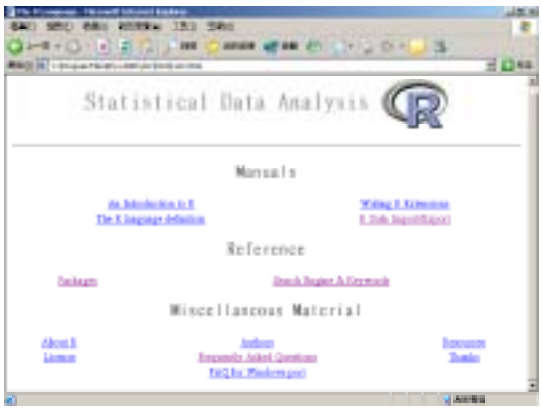


圖 2.11 線上查詢指令畫面

內容匯出成文字檔(即副檔名為.txt的檔案)或其他檔案格式(如.csv)。匯出檔案的指令為write.table, 詳細指令介紹如下：

>write.table(變數名稱, file = "檔案名稱.副檔名", quote = TRUE,

+ sep =" ",col.names = TRUE,row.names = TRUE)

其中“quote”表示是否用“" "”將變數或字元標示出來。“sep”表示欄位間的分隔符號，預設值為空隔。“row.names”和“col.names”表示是否列出行和列的名稱，預設值設定為行列的名稱皆須匯出。在此我們僅介紹常用的部分，關於其他更細部指令請自行查詢。

一般若數據不多，我們常會在R中直接輸入，但當所要分析的數據不是幾個數字，而是一個資料表的型式，那麼自行輸入似乎會有困難，除了輸入上較麻煩外，也不容易發現輸入錯誤。這時透過“匯入檔案”就會相對的簡單。所要匯入的檔案可用記事本(NotePad)先編輯完成，或較建議會使用Excel軟體的使用者，透過Excel來編輯檔案再存成文件檔，這種方法不僅介面平易近人，在輸入上也較容易。

若需要匯入的檔案已處理完成後，接下來只要下匯入檔案的指令即可將資料匯入到R中，其指令為“read.table”，輸入方式如下：

>read.table("檔案名稱.副檔名", header = FALSE, sep = " ",

+ quote = " ",row.names, col.names, na.strings = "NA")

其中“header”表示是否將檔案的第一列視為資料的變數名稱，“na.strings”表示匯入的資料若有遺失值，則以NA取代。

2.7 向量與矩陣

在2.3節我們提過可以將一長串數據指定給所設定的變數，但如何輸入這一長串數據呢？我們將在這一小節作一簡單介紹。我們可以透過向量(vector)和矩陣(matrix)的方式，將數據排列成陣列(array)或資料表(data frame)的型式。

首先，先介紹向量。向量元素(elementary)的型態可以是數字(number)，字元(character)，也可以是邏輯判斷字元(logical character)。向量的指令為“c(資料)”。三種型態的指令分別舉例如下：

1. 數字向量(numerical vector)：

```
>c(1,3,5,7,9)  ↵
[1] 1 3 5 7 9
```

2. 字元向量(character vector)：

```
>c("teacher","student")  ↵
[1] "teacher" "student"
```

3. 邏輯向量(logical vector)：

```
>c(T,T,F,T,F)  ↵
[1] TRUE TRUE FALSE TRUE FALSE
```

如何將向量指定給某一特定變數, 做法請參考2.3節, 利用“<–”。

矩陣是由向量所構成的, 元素的位置有一次序(order)關係。先看一例子。

$$x = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}。$$

利用向量的方式來表示。

```
>x<-1:12  ↵
>dim(x)<-c(3:4)  ↵
>x  ↵
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

指令中的“dim”表示變數的維度。另外亦可用較簡單的方式, 利用“矩陣”(matrix)指令。

```
>matrix(1:12,nrow=3,bycol=T)  ↵
     [,1] [,2] [,3] [,4]
```

```
[1,] 1  4  7  10
[2,] 2  5  8  11
[3,] 3  6  9  12
```

其中“nrow”表示矩陣列數, “byrow”判斷是否依照列的次序作排列。亦可透過“row.names”和“col.names”更改列和行的名稱。

在數學上, 有一常用的矩陣運算, 轉置(transformation), 表示將行跟列的位置互換。利用指令“t”即可作到轉置的運算。

其他有關矩陣的指令有“is.matrix(變數名稱)”和“as.matrix(變數名稱)”。“is.matrix”用來判斷所給定的變數是否為矩陣型式, 若是, 則會傳回“TRUE”。若傳回“FALSE”, 則為非矩陣型式。而“as.matrix”是將給定的變數轉換成矩陣型式。

2.8 使用內建程式庫

R的安裝包含有程式庫(package, 或稱套件, 函式庫), 裡面有許多相關的數據資料, 可作為練習時使用, 最基本的程式庫稱為“base”。其他的程式庫有cluster, survival, tseries等。若分析時有需要利用各種程式庫, 可透過R的首頁下載程式庫, 或者由按鈕列上程式庫(packages)的選項更新所需要的程式, 參考圖2.12。

底下我們以“base”程式庫的“airmiles”為例, 說明如何呼叫出程式庫裡的數據資料。

1. 列出程式庫裡所有檔案和簡介。參考圖2.13。指令如下：

   ```
   >data()  ↵
   ```

2. 呼叫程式庫裡某一特定資料。呼叫完成後, 利用查詢變數的指令可以看到此筆數據已經匯入。參考圖2.14。指令如下：

   ```
   >data(airmiles)  ↵
   >ls()  ↵
   ```

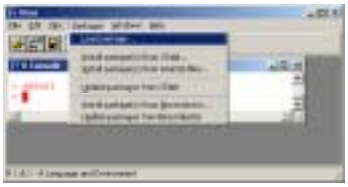


圖 2.12 查詢程式庫內容

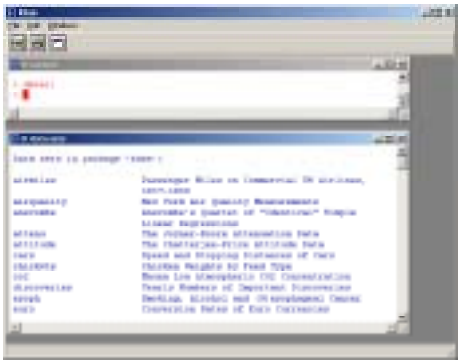


圖 2.13 查詢程式庫內容

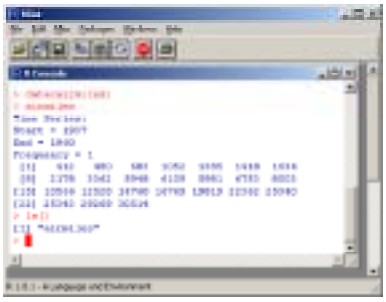


圖 2.14 呼叫程式庫中的airmiles

3. 查詢某一特定資料的詳細內容。包括此筆數據的描述(description), 使用語法(usage) , 格式(format), 詳細說明(details), 來源(source), 參考資料(references), 其他可參考的指令(see also)和例子(examples)。參考圖2.15。指令如下：

   >help("airmiles") ↵

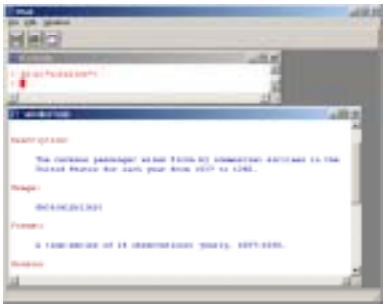


圖 2.15 查詢程式庫中airmiles的詳細內容

習　題

1. 匯入資料庫中名為“trees ”的檔案, 利用查詢指令看是否已存在。查詢完成後, 將資料匯出成名為trees的文字檔。
2. 檢查習題1中之資料是否為矩陣的型態。
3. 輸入下列矩陣：

$$X = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}。$$

第三章

統計圖

3.1 前言

從第三章開始, 將介紹統計上常用的圖形、名詞和統計方法。會針對每一章的主題說明相關名詞, 舉例說明並介紹R上的操作指令。讓使用者在讀完每一章後, 能對統計上常用的名詞及方法, 有一概略的了解, 並且知道相關的操作指令。

在獲得相關數據資料並進行分析之前, 需先對此筆資料有一粗略的了解, 才能進行各種不同的分析步驟。如果對數據沒有了解, 就進行一連串盲目的分析工作, 除了浪費時間外, 有時甚至會導致分析錯誤的情形發生。如何對手邊的資料有一全面性的了解呢？最簡單也最方便的方式便是將資料繪製成統計圖。

統計圖除了提供整筆數據的分佈情況外, 也會顯現某些趨勢, 離群值, 或者兩筆資料之間是否有何相關性存在。將數據透過圖形的展現, 不只容易觀察, 更能對整筆資料有一概略的了解。

底下我們將介紹幾個不同的統計圖。

3.2 莖葉圖

莖葉圖(stem-and-leaf diagram), 是一方便且容易畫的圖形。只要使用者選擇適當的“莖”, 並按照莖的順序依序填入資料(葉), 便可繪出數據的分佈情形。莖葉圖可畫出似直方圖的圖形, 並且保留完整數據內容, 但卻不適用於大筆的數據分析。

底下我們先以著名的老忠實噴泉(Old Faithful geyser)的資料, 來看莖葉圖的圖形樣式, 之後再針對該圖形做解釋。

此筆數據來源取自R的程式庫(做法可參考2.8節的指令), 為美國懷俄明(Wyoming)州之黃石公園(Yellowstone National Park)的老忠實噴泉的數據資料。資料內容為一資料表的型態, 共記載了兩種變數, 一為兩次噴發的時間間隔(以“waiting”表之), 一為每次噴發維持的時間(以“eruptions”表之)。部分資料內容可參考圖3.1, 共有272筆資料。

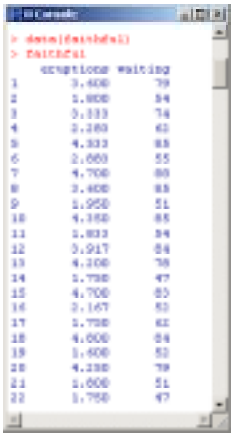


圖 3.1 老忠實噴泉數據內容

繪製莖葉圖的指令為“stem”。利用老忠實噴泉噴發的時間間隔(waiting)所得之莖葉圖如圖3.2所示。

由結果可看出, 莖葉圖的做法是先將數據做一次序的排列, 接著選擇適當的莖和葉, 並將莖和葉區隔開來。此例是以時間間隔的十位數字為莖,

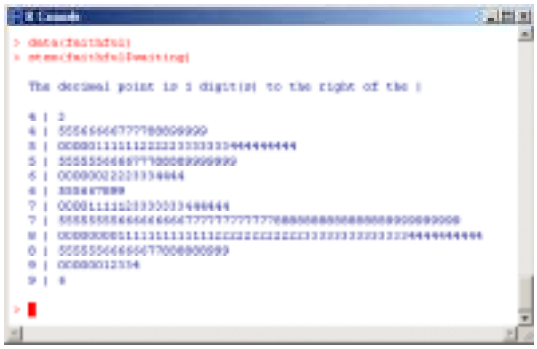


圖 3.2 老忠實噴泉數據內容

個位數字為葉, 再將兩者中間以“ | ”作區隔。可由圖形看出大部分的時間間隔皆落在75至85之間, 並且可以知道數據包含一個43, 三個45等的數據資料, 故可直接由莖葉圖的結果看出整筆數據的散佈情況。

當我們所要分析的是少數的數據時, 莖葉圖可以是一不錯的選擇, 因為它可以完整的保留數據內容。但當有一相當龐大的數據時, 選擇莖葉圖也許不是很妥當。

3.3 直方圖

直方圖(histogram), 圖中每個長方形的寬度比例於組區間, 面積則與它所代表的次數成比例。同樣以老忠實噴泉的資料分別對噴發的時間間隔和維持的時間繪製直方圖, 如圖3.3。可利用指令“hist”畫出直方圖。

3.4 長條圖

長條圖(bar plot), 是一種用寬度相同, 長度與已知次數成比例的長方形顯示數據的統計圖。可分為垂直放置或水平放置。利用“barplot”指令可畫出長條圖。

以台灣北中南(台北、台中、高雄)三區, 自1971到2000年的月平均溫度為例(資料取自中央氣象局資訊服務網站), 可得如圖3.4。

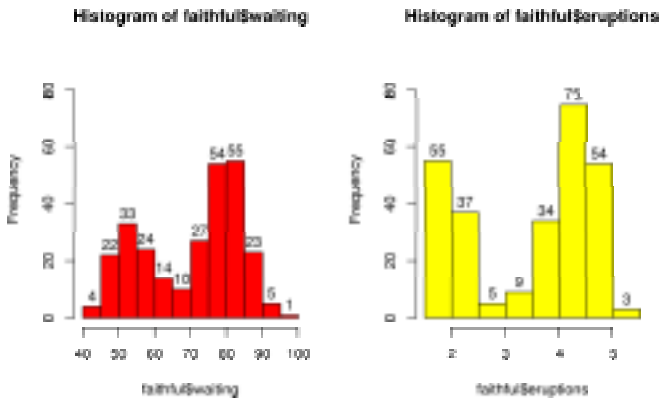


圖 3.3 老忠實噴泉的直方圖

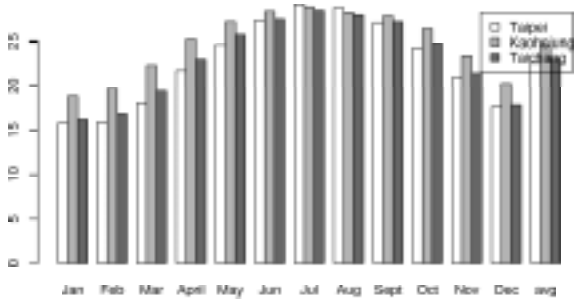


圖 3.4 台灣北中南三區域1971-2000年之月平均溫度的長條圖

另外以R程式庫中名為“Titanic ”的鐵達尼號數據資料(有關經濟地位和存活人數的列聯表)為例, 資料內容為

	是否存活	
船艙等級	死亡	存活
1st	122	203
2nd	167	118
3rd	528	178
Crew	673	212

透過此筆資料可畫出各種不同的長條圖, 如圖3.5。

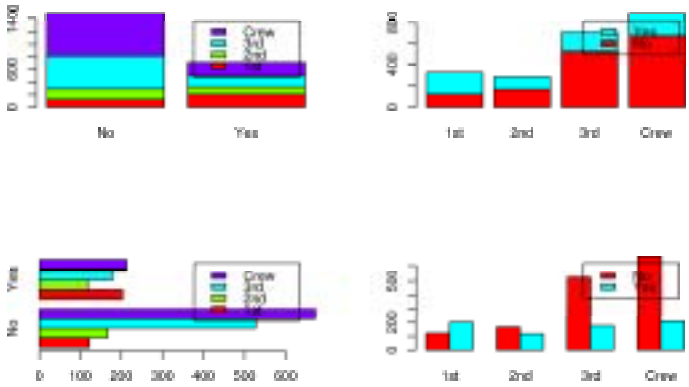


圖 3.5 關於鐵達尼號資料的不同長條圖畫法

3.5 盒鬚圖

盒鬚圖(box-and-whisker plot,簡稱boxplot)亦稱為長鬚圖, 是一很有效的表示資料的方法。可用來呈現數據分佈的某些重要特性的圖形, 也可了解資料的偏斜性(skewness)及離群值(outliers)。利用指令“boxplot”可畫出盒鬚圖。

我們以R的程式庫中名為“ToothGrowth ”的數據資料來看看盒鬚圖的圖形。此筆數據來自R的程式庫, 紀載10位實驗者牙齒長度的數據。實驗有三種不同劑量的維生素C(0.5,1,和2毫克)和兩種服用方式(柳橙汁或抗壞血酸(維生素C))。此資料共有三個變數, 長度(len), 服用方式(supp)以及劑量(dose), 共獲得60筆觀測值。

底下先透過此筆數據建立最簡單的盒鬚圖, 之後再針對不同變數來建立盒鬚圖的圖形樣式。首先若我們只對牙齒的長度有興趣, 想知道這60筆數據的分佈情況, 可透過盒鬚圖來觀察, 如圖3.5。

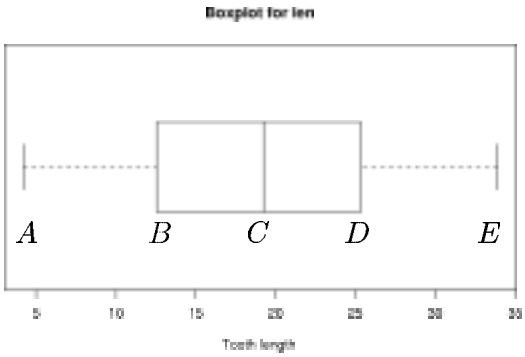


圖3.5 牙齒長度紀錄的盒鬚圖

圖3.5為一基本的盒鬚圖。圖中A為資料的最小值, E為最大值, B 和D分別為資料的上四分位數 和下四分位數, C為資料的中位數。圖中盒子包含資料的50%部分。由圖3.5可看出, 圖形包含中間的盒子, 和延伸到最大及最小值的“鬚”, 故稱為盒鬚圖。關於分位數將會在下一章有完整的介紹。

如果資料可分成數群, 則可根據不同的群繪製個別的盒鬚圖, 如圖3.6和圖3.7。

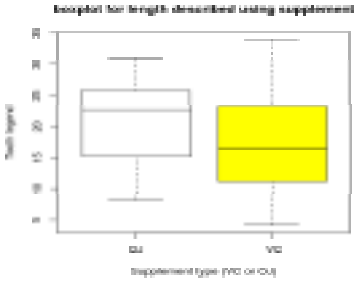


圖 3.6 資料可分為兩群的盒鬚圖

3.6 散佈圖

散佈圖(scatter plot), 乃用來說明二個變數散佈情形的直角座標圖。散佈圖上只畫數據點, 點與點之間不予連接。如果散佈圖大致落在一直線上,

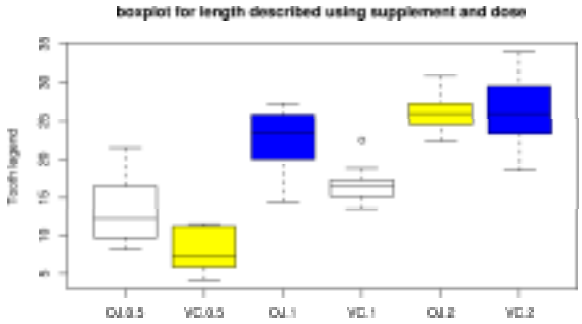


圖 3.7 多群的盒鬚圖

可能需要做相關性分析的討論。散佈圖有時也會顯示某種非線性關係(如二次曲線)、離群值等。可用指令“plot”畫出散佈圖。

我們仍以老忠實噴泉的等待時間與噴發時間為例, 畫出兩者的散佈圖形, 如圖3.8。

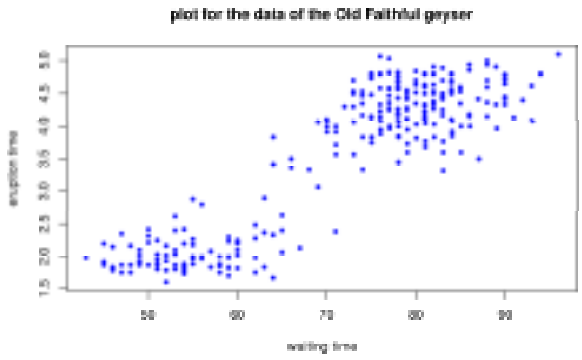


圖 3.8 老忠實噴泉資料的散佈圖

3.7 散佈圖矩陣

散佈圖矩陣(scatterplot matrices), 三變數以上資料散佈情形的關係圖。散佈圖矩陣裡的每一個圖形皆為一散佈圖。可用指令“pairs”畫出散佈圖矩陣。

我們以鳶尾(別名愛麗斯, 為一種植物)的觀測值(取自R的程式庫, 名為“iris”)為例, 來畫散佈圖矩陣。此筆數據共紀錄了50朵不同品種的鳶尾之花萼的長寬和花瓣的長寬的測量值(單位為公分)。有三種不同品種的鳶尾分別為“Iris setosa”, “versicolor”和“virginica”。圖3.9為鳶尾之散佈圖矩陣。

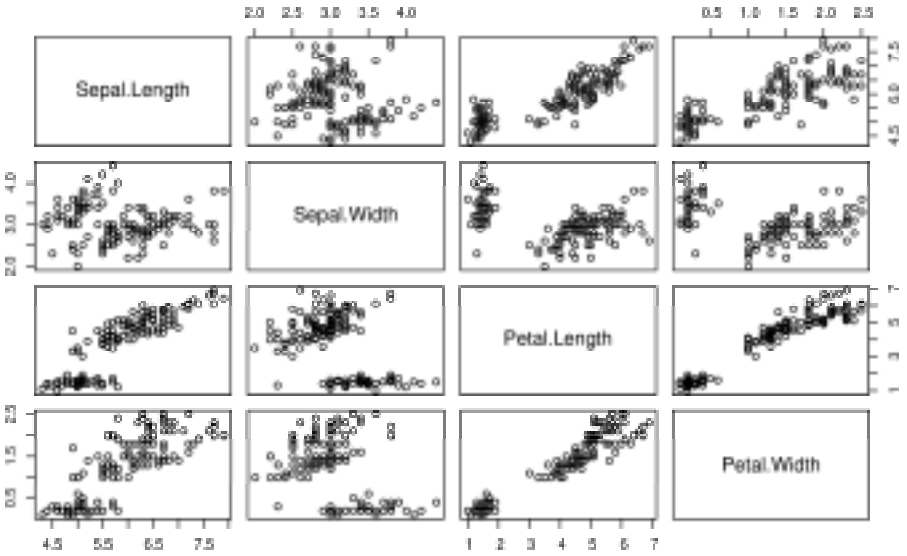


圖 3.9 鳶尾的散佈圖矩陣

3.8 常態分位圖

分位圖(Quantile-Quantile plot, Q-Q plot), 利用樣本分位數來描繪資料的圖。以圖形來檢驗數據是否來自某一分佈的方法。

常態分佈的Q-Q圖(Normal Quantile-Quantile plot , Normal Q-Q plot), 在實際的應用裡, 常會想知道收集到的資料是否來自常態分佈(參考第五章)的母體, 可利用常態分佈的Q-Q 圖來做初步的檢驗。這是一種簡易檢驗常態性(normality)的方法。若樣本是來自常態分佈的母體, 則其Q-Q 圖近似於一條直線, 否則便不是。可用指令“qqnorm ”畫出常態分佈的Q-Q圖。

底下我們利用隨機亂數(random number , 常用來選擇樣本, 或用來作為模擬的研究)來產生所需之數據, 並繪製其常態分佈之Q-Q圖。我們利用t分佈來產生一組隨機亂數, 利用圖形來判斷是否來自常態分佈。

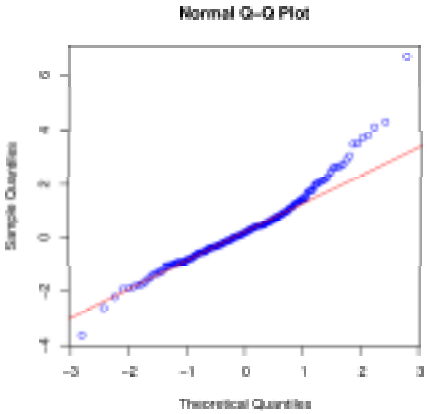


圖 3.10 自由度為4的t分佈的常態分佈Q-Q圖

圖3.10是利用自由度為4的t分佈生成的數據資料, 所畫出的常態分佈之Q-Q圖。可看出圖形右上方有偏離的情形。遇到這種情況就該察覺是否該針對資料做進一步的檢查, 以確定資料是否來自常態分佈。此處因我們是以t分佈來產生隨機亂數, 故確實資料並不是來自常態分佈。若所獲得的資料並不能確定來自哪一分佈, 則只要出現如圖3.10的情況, 就要提醒自己是否有錯誤產生, 如輸入錯誤, 或者是資料有離群值存在等問題。

上面的例子是以自由度為4的t分佈所畫出來的圖形。我們知道自由度為30的t分佈便近似常態分佈。由圖3.11的圖形可看出，常態分佈之Q-Q圖近似一條直線。故由模擬的資料也可以知道自由度為30的t分佈近似常態分佈。

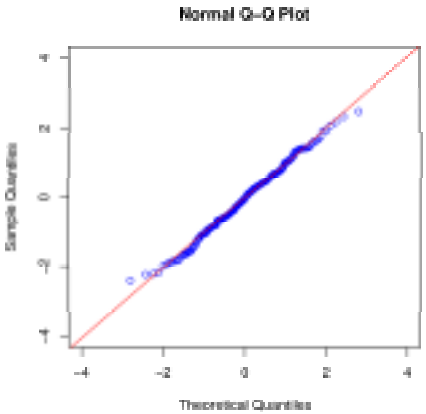


圖 3.11 自由度為30的t分佈之常態分佈Q-Q圖

3.9 時間序列圖

時間序列(time series)，乃依照時間先後順序所紀錄的一組觀測值。

時間序列圖(time series plots)，為一種用來觀測時間序列資料的方式。透過時間序列圖可觀察出資料是否有某些趨勢，如季節效應等，對於時間序列的分析有很大的幫助。可用指令“plot ”或“ts.plot ”畫出時間序列圖。

底下以太陽黑子的資料(取自R的程式庫，名為sonspots)為例，畫出其時間序列圖。資料內容為自1749到1983年所觀測到太陽黑子活動的月平均值。所得圖形如圖3.12。

相關的研究指出太陽黑子的活動有十一年的週期變化。由時間序列圖同樣可看出圖形有一週期的變化。詳細的分析還得做進一步的探討，在此

並不作深入的討論, 若讀者有興趣可參考時間序列的相關書籍。

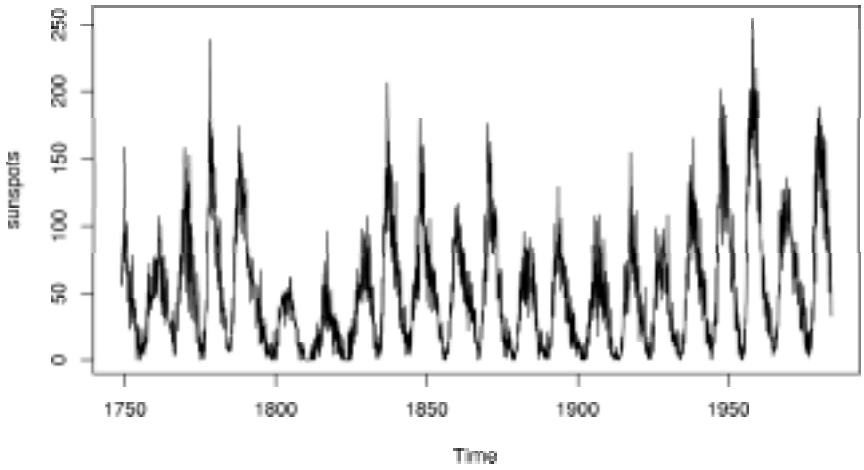


圖 3.12 太陽黑子的時間序列圖

3.10 使用統計軟體R

3.2節 莖葉圖指令說明

繪製莖葉圖的指令為“stem (變數名稱)”。畫出3.2節圖形的指令步驟為：

1. 輸入老忠實噴泉的資料。

   > data(faithful)

2. 畫出老忠實噴泉等待時間的莖葉圖。

   > stem(faithful$waiting)

3.3節 直方圖指令說明

繪製直方圖的指令為“hist (變數名稱)”。畫出3.3節圖形的指令步驟為：

1. 分割圖形視窗。利用par 指令將圖形分成左右兩個位置, mfrow 表示分割是依據列來做切割, 若改為mfcol 則根據行來做切割。

   > par(mfrow=c(1,2))

2. 畫出老忠實噴泉等待時間的直方圖。

   > hist(faithful$waiting,freq=T,col="red",

   + labels=T,ylim=c(0,90))

   > hist(faithful$eruptions,freq=T,col="yellow",

   + labels=T,ylim=c(0,90))

在繪製直方圖的指令中, 可看到“freq =T”, 表示依據資料的頻率來繪製圖形。若輸入“freq=F”, 則會依據密度(density)來繪製圖表。圖形填充顏色的指令可用“col ”, 直接輸入顏色的英文名字, 如紅色(red), 藍色(blue)等, 或者利用“灰階”(數字越大表示越靠近白色, 如gray50, gray80)來當填充的顏色。“lables ”表示是否標示出直方圖次數。“xlim ”,

“ylim ”分別代表x軸和y軸的長度和高度。另外可增加“breaks ”來設定直方圖的長度, breaks可等於單一值, 一向量值, 或以一函數來計算每一直方圖的次數並做分割。

3.4節 長條圖指令說明

繪製直方圖的指令為

> barplot (變數名稱, beside=FALSE, horiz=FALSE, legend=NULL)

其中“beside ”表示長條圖是每一類別皆成一獨立的長條圖, 或者堆疊成單一長條圖。“horiz ”設定圖形水平放置或垂直放置。“legend ”表示是否標示出每一類長條圖代表的意義。

畫出3.4節台灣北中南平均溫度圖形的指令步驟為：

1. 輸入台灣北中南平均溫度資料。

   > areatemp<-read.table("temp_aveg.txt",header=T)

2. 將資料檔案轉成矩陣格式。因資料必須為向量或矩陣, 故利用“as.matrix ”對資料作矩陣的轉換。

   > areatemp<-as.matrix(areatemp)

3. 判斷資料是否為矩陣格式。此步驟可省略。

   > is.matrix (areatemp)

4. 畫平均溫度的長條圖。

   > barplot(areatemp,col=c("white", "gray70","gray40"),beside=T,

   + legend=rownames(areatemp))

畫出3.4節鐵達尼號經濟地位與存活人數關係的長條圖的指令步驟為：

1. 將繪圖視窗分隔成四個區塊。

   > par(mfrow=c(2,2))

2. 輸入鐵達尼號數據資料。

   > data(Titanic)

3. 將資料整理成為以經濟地位和是否存活人數的列聯表。

   > SurTitanic<-apply(Titanic,c(1,4),sum)

4. 畫出以存活與否為分類的長條圖。將同一類別的資料堆疊成單一長條圖, 圖中並不特別標明個數值。

   > barplot(SurTitanic,beside=F,legend=T,col=rainbow(4))

5. 畫出以經濟地位為分類的長條圖。將同一類別的資料堆疊成單一長條圖, 圖中並不特別標明個數值。

   > barplot(t(SurTitanic),beside=F,legend=T,col=rainbow(2))

6. 畫出以存活與否為分類的長條圖。水平放置, 並標明個數值。

   > barplot(SurTitanic,beside=T,legend=T,

   + col=rainbow(4),horiz=T)

7. 畫出以經濟地位為分類的長條圖。垂直放置, 並標明個數值。

   > barplot(t(SurTitanic),beside=T,legend=T,col=rainbow(2))

3.5節 盒鬚圖指令說明

繪製盒鬚圖的指令為

> boxplot (變數名稱, boxwex=0.8, col=NULL, horizontal=FALSE,

+ add=FALSE, at=NULL)。

其中“boxwex ”為圖中盒子的寬度。“add ”使否將圖形加至前一次的繪圖視窗中, 或者是覆蓋前一次的圖形。“at ”設定盒鬚圖所要放置的位置。

畫出3.5節牙齒生長情形的指令步驟為：

1. 輸入牙齒生長數據資料。

```
> data(ToothGrowth)
```

2. 畫牙齒生長的長度變數的盒鬚圖。

```
> boxplot(ToothGrowth$len,col="blue",main="Boxplot for len",
+ xlab="Tooth length",horizontal=T)
```

如果資料可分成數群, 則可根據不同的群繪製個別的盒鬚圖, 可有多種畫法, 底下為兩種較簡單的畫法。

1. 根據變數supp的類別, 將變數len畫成不同群的盒鬚圖。

```
> boxplot(ToothGrowth$len~ToothGrowth$supp, col=c("white","yellow"),
+ main="boxplot fo length described using supplement",
+ ylab="Tooth legend",xlab="Supplement type (VC or OJ)")
```

2. 根據變數supp和變數dose的類別組合, 將變數len畫成不同群的盒鬚圖。

```
> boxplot(len~supp*dose,data=ToothGrowth,
+ col=c("white","yellow","blue") ,ylab="Tooth legend",
+ main="boxplot for length described using supplement and dose")
```

指令中的“ToothGrowth$len~ToothGrowth$supp”表示“ToothGrowth $len”是根據“ToothGrowth$supp”來分群, 另外“main ”, “xlab ”和“ylab ”分別表示整個圖形的標題, x軸標籤和y軸標籤。

3.6節 散佈圖指令說明

繪製散佈圖的指令為

```
>plot(x, y, xlim=range(x), ylim=range(y), type="p",
+main, xlab, ylab, pch="1", col="blue")
```

指令中“type ”表示圖形的樣式, 圖形的樣式指令有p, l, b, h, s, 分別表示點圖, 線段圖, 點線圖, 長條圖, 階梯圖。“pch ”代表點的樣式, 所有點的樣式可參考圖3.13。

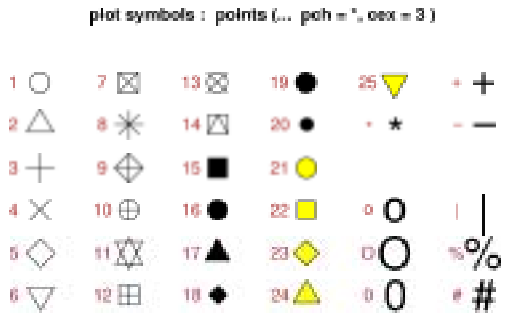


圖 3.13 圖形標示的點的所有樣式

畫出3.6節老忠實噴泉散佈圖的指令步驟為：

1. 畫出老忠實噴泉等待時間和噴發時間的散佈圖。

   > plot(faithful$waiting,faithful$eruptions,

   + type="p",main="plot for the data of the Old Faithful geyser",

   + xlab="waiting time",ylab="eruption time",pch=20,col="blue")

3.7節 散佈圖矩陣圖指令說明

散佈圖矩陣指令為

>pairs (變數名稱)

指令中的變數需為矩陣物件。

畫出3.7節鳶尾散佈圖矩陣的指令步驟為：

1. 輸入鳶尾的數據資料。

   >data(iris)

2. 畫出散佈圖矩陣。

```
>pairs(iris)
```

若只想畫出前四個變數的散佈圖矩陣, 可下“pairs(iris[1:4])”指令。

3.8節 常態分位圖指令說明

畫出常態分位圖指令為

```
>qqnorm (變數名稱)
>qqline (變數名稱)
```

指令“qqnorm”是產生常態分佈的Q-Q圖的函數指令。“qqline”表示在常態分佈的Q-Q圖中加入通過第一和第三分位數的直線。可利用此直線來判斷圖形是否近似一條直線。

利用t分佈產生隨機亂數, 並利用此筆資料畫出常態分位圖的指令步驟為：

1. 產生200筆自由度為4的t分佈的隨機亂數資料。

   ```
   > rn<-rt (200,df=4)
   ```

2. 畫出常態分位圖。

   ```
   > qqnorm(rn,col="blue")
   ```

3. 在常態分佈的Q-Q圖中加入通過第一和第三分位數的直線。

   ```
   > qqline(rn,col="red")
   ```

3.9節 時間序列圖指令說明

時間序列圖操作指令為

```
>plot(變數名稱)
```

與前面所提過的散佈圖指令是相同的, 唯一不同點為變數須為時間序列物件。

畫出3.9節太陽黑子時間序列圖的指令步驟為：

1. 從程式庫輸入太陽黑子資料。

   > data(sunspots)

2. 畫出太陽黑子的時間序列圖。因為數據資料為時間序列的物件, 所以所得之圖形為時間序列圖。

   > plot(sunspots, main = "sunspots data", xlab = "Year",

   + ylab= "Monthly sunspot numbers")

習　題

1. 畫出資料庫中名為“trees”的盒鬚圖及散佈圖矩陣。
2. 同上題之數據, 利用矩陣轉置的指令並搭配長條圖的指令, 畫出其長條圖, 並標示出每一長方條所代表的類別。
3. 生成一組樣本數為100且其自由度為10之t分佈的數據資料, 並繪出常態分佈的Q-Q圖。

第四章

敘述統計量

4.1 前言

在一組資料中, 有時我們會想了解, 這組樣本中的最大值, 最小值或者是中位數。可以透過資料簡化的過程, 幫助我們更進一步了解此筆資料所攜帶的資訊。

4.2 最大和最小值

對於不同的數據資料, 有些情況會想要知道此筆數據的最小值或最大值。例如, 近幾年來全球的最高溫度, 一年中翡翠水庫(Feitsui Reservoir)的最低水位。

將一組資料, 按照順序由小至大排好, 則整筆資料中最小的值就是所謂最小值(minimun value), 而最大的值就稱為最大值(maximum value)。另外, 最大值和最小值的差距稱為全距(range)。

在R中, 可用"min", "max"和"range"指令分別計算出整筆資料的最小值, 最大值和全距。

舉例來說, 臺北縣新店市的翡翠水庫92年全年最低水位為133.5公尺, 最高水位為161.52公尺。

資料來源：台北翡翠水庫管理局網站(http: // www.feitsui.gov.tw/)。

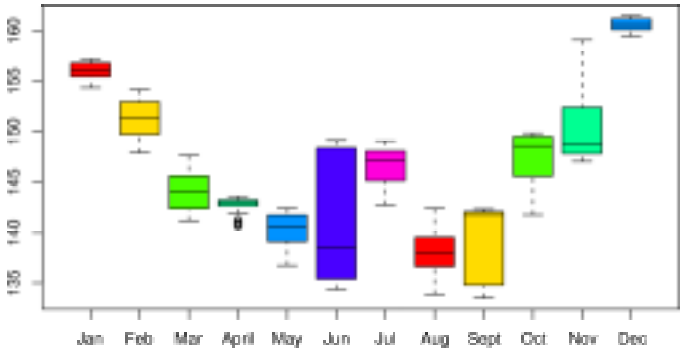


圖 4.1 翡翠水庫各月份日平均水位的盒鬚圖

4.3 中位數、分位數和平均值

在美國, 職業球隊裡, 例如職棒, 極少數的球員薪水很高, 但是大部分的球員薪水是很低的。若僅單純由平均值(mean)來看, 是無法獲得此訊息的。當討論到收入、價格等含有一些較極端的值(過高或過低), 中位數(median)可能是一較合理的指標。

中位數是一組資料中處於中間位置之值, 所有比它小和比它大的數具有相等的次數或機率。若資料數n為奇數, 則中位數為第$\frac{n+1}{2}$ 位之值。若為偶數, 則中位數為第$\frac{n}{2}$位和第$\frac{n+1}{2}$位之值的平均。其指令為“median”。

與中位數有關的一個統計量為“分位數”。在此利用第p分位數(pth quantile)來介紹分位數。第p分位數指的是某一點x_p, 對比x_p小於或等於的值之機率大於或等於p, 及對比x_p大於或等於的值之機率大於或等於$1-p$, 則x_p稱為第p分位數。而第25分位數稱為下四分位數, 第50分位數稱為中

位數, 第75分位數稱為上四分位數 。我們可用“quantile”指令求出資料的第0, 25, 50, 75, 100分位數。

另外一個跟分位數有關的值是“內四分位數間距”(interquantile range)。內四分位數間距即表示一組資料的上四分位數與下四分位數之差。指令為“IQR”。

在前面有提到關於“平均值”, 指的是樣本平均, 表示將資料的和除以總個數, 亦即$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$。平均值的指令為“mean”。

根據翡翠水庫92年的日平均水位數據, 全年的中位數為145.73公尺, 全年日平均水位的平均值為146.46公尺。翡翠水庫的日平均水位的第0, 25, 50, 75, 100分位數, 如表4.2。

表4.1 翡翠水庫日平均水位的各月份平均值

月份	一月	二月	三月	四月	五月	六月
平均水位(公尺)	156.02	151.28	144.09	142.72	140.29	141.08
月份	七月	八月	九月	十月	十一月	十二月
平均水位(公尺)	146.60	138.05	139.26	147.14	150.48	160.64

表4.2 翡翠水庫的日平均水位的分位數資料

分位數	0%	25%	50%	75%	100%
日平均水位(公尺)	133.50	141.69	145.73	150.62	161.52

4.4 變異數和標準差

有些時候我們會有多筆的資料需要同時分析, 但光由資料的平均值來代表整筆數據, 可能無法了解這些資料彼此之間的差異情況。例如有兩組學生同時參加一測驗, 每組各有三人。平均成績皆為60分, 但第一組之成績

為61, 60, 59；第二組之成績為90, 60, 30。顯然第一組學生的程度相當接近, 而第二組學生的程度則有很大的差異。由此例我們可以看到, 除了平均值, 變異數(variance)能提供關於資料散佈情況的資訊, 有助於了解資料的概況。

變異數為量測一組數據$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$離散程度的一種方式。在統計學上, 以$\sigma^2$代表變異數。在此我們探討的是"樣本變異數"(sample variance), 以s^2表之, 即

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n-1},$$

其中$\bar{\boldsymbol{x}}$表示樣本平均值。樣本變異數的指令為"var"。

另一個量測數據離散或散佈程度的方式為標準差(standard deviation)。在統計學上, 以σ來表示。變異數的正平方根即稱為標準差。在此我們仍是討論樣本標準差, 以s表之, 亦即

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n-1}}。$$

若有兩組數據$\boldsymbol{x}$, $\boldsymbol{y}$則分別以$s_{\boldsymbol{x}}$, $s_{\boldsymbol{y}}$來表示。標準差的指令為"sd"。

表4.3列出翡翠水庫日平均水位之各月份的變異數。由表4.2可知六月份平均水位為141.08公尺, 並非最低值, 表4.3顯示出六月份日平均水位的變異數卻達到38.0063, 為各月份最高值, 可知六月份的日平均水位改變相當大, 此關係亦可由圖4.1的盒鬚圖大略看出。

表4.3 翡翠水庫日平均水位各月份的變異數

月份	一月	二月	三月	四月	五月	六月
平均水位(公尺)	0.6940	3.6208	4.0190	0.6421	2.8483	38.0063
月份	七月	八月	九月	十月	十一月	十二月
平均水位(公尺)	3.6634	5.1584	12.8414	8.6029	11.2563	0.4043

4.5 共變異數

若有二組資料, 此兩組資料可能會互相影響, 也可能不會。如果二組資料會互相有影響, 則二資料間存在某種關係。可透過共變異數(covariance)來討論二組資料間線性關係的強弱。較大且為正的共變異數, 顯示二組資料, 大致以線性的方式同時增大或減小；較大且為負的共變異數, 顯示二組資料, 大致以線性的方式一增大時另一減小。共變異數很接近0時, 表二組資料的線性關係很弱。

共變異數為二組資料$\boldsymbol{x}$, $\boldsymbol{y}$的一個統計量, 以$s_{\boldsymbol{xy}}$來表示。樣本共變異數的計算式為

$$s_{\boldsymbol{xy}} = \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{n},$$

其中$\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{y}}$分別表示樣本$\boldsymbol{x}$和$\boldsymbol{y}$的平均值。共變異數的指令為“cov”。

以老忠實噴泉的數據資料為例, 觀察等待時間與噴發時間是否有何關係存在。可算得兩者的共變異數為13.9778, 可知大致以線性的方式同時增大。

4.6 相關係數

相關係數(correlation coefficient, 或只稱correlation)也可用來度量二組資料的關係強弱。以共變異數來量測二組資料$\boldsymbol{x}$及$\boldsymbol{y}$的變化情況有一缺點, 就是其值與量測$\boldsymbol{x}$, $\boldsymbol{y}$ 的尺度有關。透過相關係數便可以消除這種困擾, 它永遠取值在區間$[-1, 1]$。相關係數等於1或-1時, 表示$\boldsymbol{x}$與$\boldsymbol{y}$有完美的線性關係。相關係數的計算方式為

$$\mathrm{Cor}(\boldsymbol{x}, \boldsymbol{y}) = \frac{s_{\boldsymbol{xy}}}{s_{\boldsymbol{x}} s_{\boldsymbol{y}}}。$$

相關係數的指令為“cor”。

在前一小節, 針對老忠實噴泉的等待時間與噴發時間計算其共變異數, 為13.9778, 只能判斷兩者大致以線性的方式同時增大。但是關係強弱與否

卻無從得知, 可利用相關係數來了解其關係的強弱, 我們計算得兩者的相關係數為0.9008。可判斷兩者有很強的線性關係。

4.7 使用統計軟體R

4.2節 最大和最小值

1. 輸入翡翠水庫92年日平均水位數據資料。

   > feitsui<−read.table("feitsui_ month.csv",header=T, ", sep=",",

   + na.strings=".")

2. 計算翡翠水庫92年日平均水位最大值。

   > max(feitsui,na.rm=FALSE)

   其中的“na.rm”表示是否忽略遺失值。

3. 計算翡翠水庫92年日平均水位最小值。

   > min(feitsui,na.rm=FALSE)

4. 畫出翡翠水庫92年日平均水位之月資料盒鬚圖。

   > boxplot(feitsui,col=rainbow(7))

   其中顏色的指令“col”可直接利用彩虹(rainbow)的顏色來設定。

4.3節 中位數, 分位數和平均值

1. 計算翡翠水庫92年日平均水位之各月的平均值。

   > mean(feitsui,na.rm=T)

2. 計算翡翠水庫92年日平均水位全年的平均值。

   > sum(feitsui,na.rm=T)/365

3. 計算翡翠水庫92年日平均水位全年的中位數。

   利用“as.matrix”將資料轉成矩陣型式, 再計算全年的中位數。

   > feitsui<−as.matrix(feitsui)

```
> median(feitsui,na.rm=T)
```

另外可利用“apply”指令計算出翡翠水庫92年日平均水位之各月份的中位數, 指令如下

```
> apply(feitsui, 2, median,na.rm=T)
```

4. 計算翡翠水庫92年日平均水位全年的第0, 25, 50, 75, 100分位數。

```
> quantile(feitsui,na.rm=T)
```

4.4節 變異數和標準差

1. 計算出翡翠水庫92年日平均水位之各月份的變異數。

```
> apply(feitsui, 2, var,na.rm=T)
```

4.5節 共變異數

1. 計算出老忠實噴泉的等待時間與噴發時間的共變異數。

```
> cov(faithful$waiting,faithful$eruptions)
```

4.6節 相關係數

1. 計算出老忠實噴泉的等待時間與噴發間隔的相關係數。

```
> cor(faithful$waiting,faithful$eruptions)
```

習　題

1. 利用下列資料計算出本章所提之各敘述統計量值。

資料一：92, 34, 80, 76, 80, 67, 86, 93, 20, 98。

資料二：60, 45, 50, 81, 14, 29, 90, 41, 47, 15。

第五章

機率分佈

5.1 前言

在本章中, 將介紹一些機率上常見的分佈, 並作一簡單的介紹和說明。若欲進一步閱讀請參考黃文璋(2003)。

5.2 二項分佈

對一可能只有兩種結果的隨機試驗, 便稱為一伯努力試驗(Bernoulli trial)。而一隨機變數X若滿足

$$P(X=0)=1-p, \quad P(X=1)=p,$$

其中$0 \le p \le 1$, 便稱為具有參數p之伯努力分佈, 以$\mathcal{B}er(p)$表之。

觀測n個獨立的伯努力試驗, 每次成功之機率皆設為p, 又以X表總共成功之次數, 則X之機率密度函數(簡稱p.d.f.)可表為

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x}, \quad x=0,1,\cdots,n。 \tag{5.1}$$

一隨機變數之p.d.f.若有如(5.1)之型式, 便稱為有參數n及p之二項分佈(binomial distribution), 以$\mathcal{B}(n,p)$表之, 其中n為一正整數, $0 \le p \le 1$。利

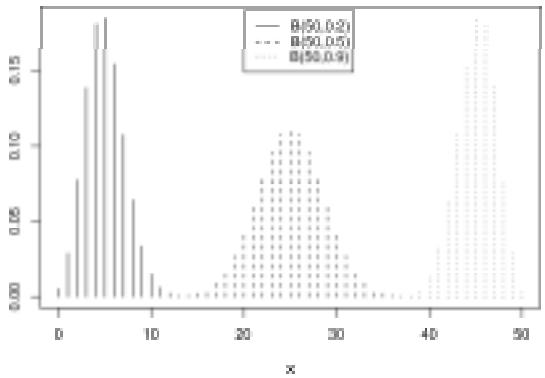


圖 5.1 $\mathcal{B}(50,0.1)$, $\mathcal{B}(50,0.5)$, $\mathcal{B}(50,0.9)$分佈之p.d.f.的長條圖

用二項分佈, 可知投擲一公正銅板($p=0.5$)50次, 出現20次正面的機率為0.041859。圖5.1列出一些二項分佈p.d.f.的圖形。

5.3 幾何分佈

持續地投擲一出現正面的機率為$p>0$之銅板, 每次投擲設為獨立, 若令X表總共須投擲的次數, 則

$$P(X=x)=p(1-p)^{x-1}, x=1,2,\cdots。 \tag{5.2}$$

一p.d.f.如(5.2)式之隨機變數, 便稱為有參數p之幾何分佈(geometric distribution), 以$\mathcal{G}e(p)$表之。(5.2)式定義出一幾何級數, 這是此分佈之由來。幾何分佈就是進行一數列之獨立的伯努力試驗, 但次數並未事先固定, 而是直至成功一次才停止。圖5.2給出$\mathcal{G}e(0.1)$及$\mathcal{G}e(0.5)$p.d.f.之圖形。

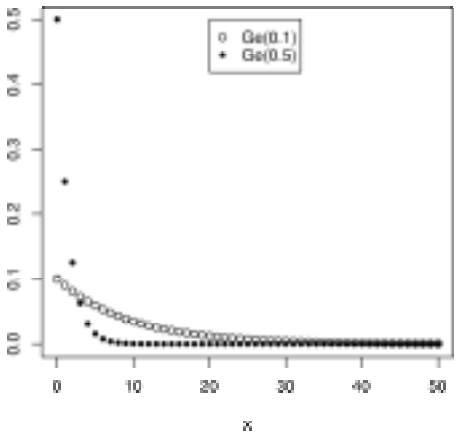


圖 5.2 $\mathcal{G}e(0.1)$及$\mathcal{G}e(0.5)$p.d.f.之圖形

5.4 波松分佈

一隨機變數X, 若其p.d.f.為

$$f(x) = \frac{e^{\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, \cdots,$$

其中$\lambda > 0$, 便稱為有參數λ之波松分佈(Poisson distribution), 以$\mathcal{P}(\lambda)$表之。圖5.3給出三個波松分佈p.d.f.之圖形。

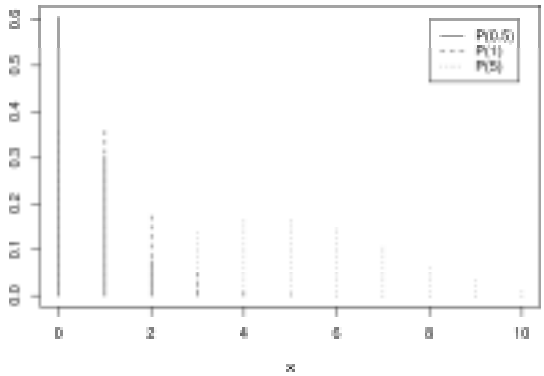


圖 5.4 $\mathcal{P}(0.5)$, $\mathcal{P}(1)$, $\mathcal{P}(5)$p.d.f.之圖形

5.5 Gamma 分佈

隨機變數X, 若其p.d.f.為

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad x > 0, \tag{5.3}$$

其中α, $\beta > 0$為二正數, 便稱為有參數α, β之Gamma 分佈(gamma distribution), 以$\Gamma(\alpha, \beta)$表之。Gamma分佈中的參數α稱為形狀參數(shape parameter), β稱為尺度參數(scale paremeter)。此因α影響p.d.f.圖形之陡峭程度, 而β影響散佈程度。圖5.4給出一些Gamma分佈之圖形。

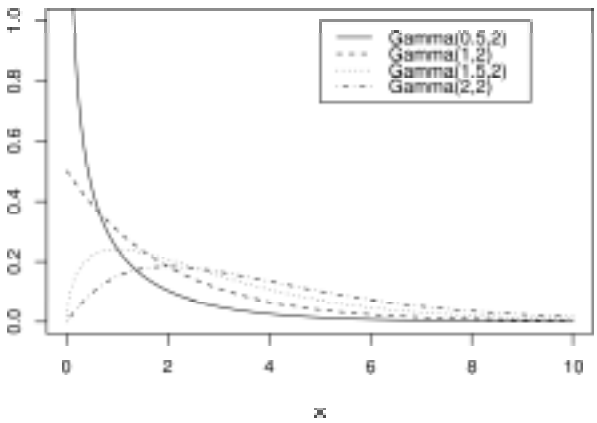


圖 5.3 $\Gamma(0.5, 2)$, $\Gamma(1, 2)$, $\Gamma(1.5, 2)$, $\Gamma(2.5, 2)$分佈p.d.f.之圖形

5.6 卡方分佈

如果Gamma分佈之$\alpha = p/2$, 其中p為一正整數, $\beta = 2$, 便得到卡方分佈(chi-squared distribution)。卡方分佈只有一參數p, 稱為其自由度(degrees of freedom), 而p.d.f.為

$$f(x) = \frac{x^{p/2-1}e^{-x/2}}{\Gamma(p/2)2^{p/2}}, \quad x > 0。$$

此分佈以χ_p^2表之。圖5.4中之Gamma分佈p.d.f.之圖形, 由於對應$\beta = 2$, 故皆為卡方分佈p.d.f.之圖形, 分別為χ_1^2, χ_2^2, χ_3^2及χ_4^2。

5.7 常態分佈

常態分佈(normal distribution)又稱高斯分佈。常態分佈有二參數μ及σ^2, $\mu \in R, \sigma > 0$, 以$\mathcal{N}(\mu, \sigma^2)$表此分佈, p.d.f.為

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in R,$$

其中μ稱為位置參數(location parameter), 此因μ影響圖形的位置(如圖5.5), σ則為尺度參數(參考圖5.6)。

常態分佈的圖形為鐘形曲線(bell-shaped curve), 且具有對稱性。常態分佈很適合當作不少母體的機率模式。

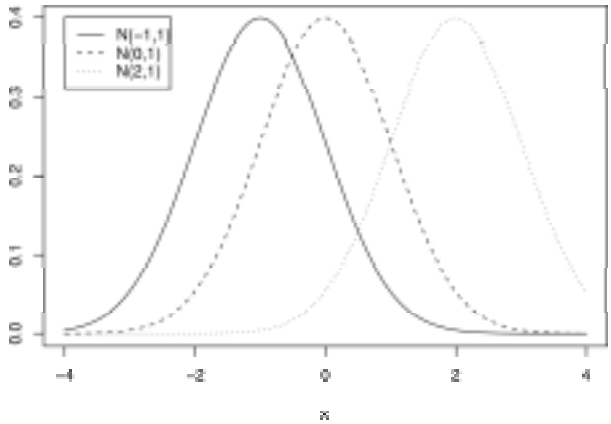


圖 5.5 $\mathcal{N}(-1,1), \mathcal{N}(0,1), \mathcal{N}(2,1)$p.d.f.之圖形

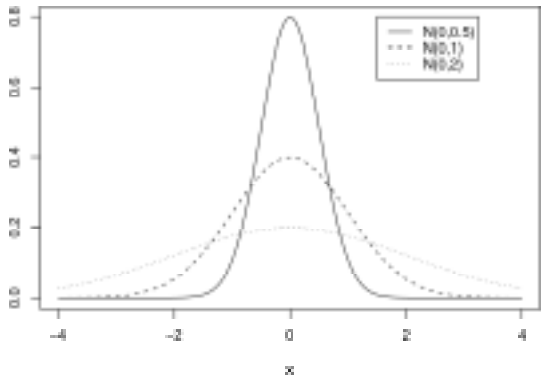


圖 5.6 $\mathcal{N}(0,0.5), \mathcal{N}(0,1), \mathcal{N}(0,2)$p.d.f.之圖形

5.8 Beta 分佈

Beta分佈(beta distribution)定義在區間$(0,1)$，有二參數α, $\beta > 0$, 以$\mathcal{B}e(\alpha,\beta)$表之, 此分佈的p.d.f.為

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1。 \tag{5.4}$$

Beta分佈是少數取值在一有限區間的常見分佈, 可用來當作取值在0至1的母體之機率模式。由於有二參數, 參數的改變, 可使圖形有很大的變化(見圖5.7)。

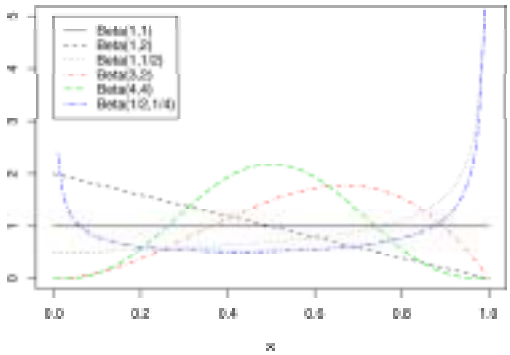


圖 5.7 不同Beta分佈的p.d.f.圖形

5.9 t 分佈

設X, Y為二獨立的隨機變數, 分別有$\mathcal{N}(0,1)$及χ_n^2分佈。令

$$T = \frac{X}{\sqrt{Y/n}},$$

則T稱有自由度n之t 分佈(t distribution), 以$\mathcal{T}_n$表之。

設$X_1, \cdots, X_n$為一組由$\mathcal{N}(\mu, \sigma^2)$分佈所產生之隨機樣本。則

$$T = \frac{\overline{X} - \mu}{(S_n/\sqrt{n})}$$

有$\mathcal{T}_{n-1}$分佈, 其中$\overline{X}$和S_n分別代表樣本平均和樣本標準差。圖5.8給出幾個t分佈p.d.f.的圖形。

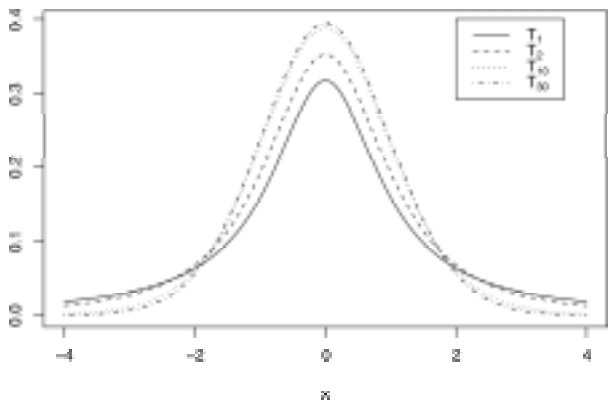


圖 5.8 不同t分佈的p.d.f.圖形

5.10 $\mathcal{F}$ 分佈

設U, V為二獨立的隨機變數, 分別有χ_n^2及χ_m^2分佈。令

$$F = \frac{U/n}{V/m},$$

則稱F有自由度n及m之$\mathcal{F}$分佈($\mathcal{F}$ distribution)。圖5.9給出幾個$\mathcal{F}$分佈p.d.f.的圖形。

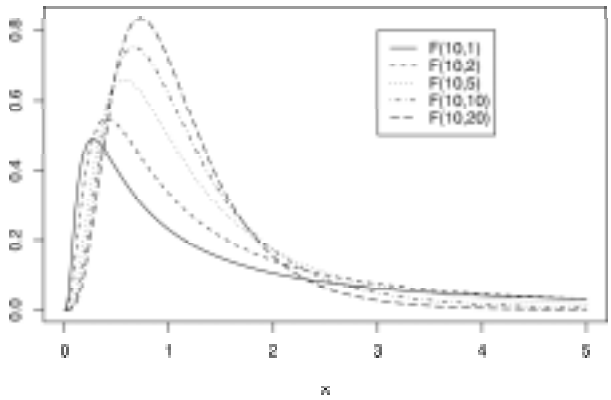


圖 5.9 不同$\mathcal{F}$分佈的p.d.f.圖形

5.11 分佈指令表

表5.1 機率上常見分佈在R的指令表

分佈名稱	指令	說明
二項分佈 $\mathcal{B}$(size,prob)	dbinorm(x,size,prob)	二項分佈中x所對應的密度
	pbinormm(q,size,prob)	二項分佈中q所對應的累積機率
	qbinorm(p,size,prob)	二項分佈中機率值為p所對應的分位點
	rbinorm(n,size,prob)	利用二項分佈產生n個隨機亂數
幾何分佈 $\mathcal{G}$e(prob)	dgeom(x,prob)	幾何分佈中x所對應的密度
	pgeom(q,prob)	幾何分佈中q所對應的累積機率
	qgeom(p,prob)	幾何分佈中機率值為p所對應的分位點
	rgeom(n,prob)	利用幾何分佈產生n個隨機亂數
波松分佈 $\mathcal{P}$(lambda)	dpois(x,lambda)	波松分佈中x所對應的密度
	ppois(q,lambda)	波松分佈中q所對應的累積機率
	qpois(p,lambda)	波松分佈中機率值為p所對應的分位點
	rpois(n,lambda)	利用波松分佈產生n個隨機亂數
Gamma分佈 Γ(alpha,beta)	dgamma(x,alpha,1/beta)	Gamma分佈中x所對應的密度
	pgamma(q,alpha,1/beta)	Gamma分佈中q所對應的累積機率
	qgamma(p,alpha,1/beta)	Gamma分佈中機率值為p所對應的分位點
	rgamma(n,alpha,1/beta)	利用Gamma分佈產生n個隨機亂數
卡方分佈 χ^2_{df}	dchisq(x,df,ncp=0)	卡方分佈中x所對應的密度
	pchisq(q,df,ncp=0)	卡方分佈中q所對應的累積機率
	qchisq(p,df,ncp=0)	卡方分佈中機率值為p所對應的分位點
	rchisq(n,df,ncp=0)	利用卡方分佈產生n個隨機亂數
常態分佈 $\mathcal{N}$(mean,sd^2)	dnorm(x,mean,sd)	常態分佈中x所對應的密度
	pnorm(q,mean,sd)	常態分佈中q所對應的累積機率
	qnorm(p,mean,sd)	常態分佈中機率值為p所對應的分位點
	rnorm(n,mean,sd)	利用常態分佈產生n個隨機亂數

*指令中的“ncp”代表非中心參數(non-centrality parameter), 預設值為0。

表5.1 機率上常見分佈在R的指令表(續)

分佈名稱	指令	說明
Beta分佈 $\mathcal{B}$eta(alpha,beta)	dbeta(x,alpha,beta) pbeta(q,alpha,beta) qbeta(p,alpha,beta) rbeta(n,alpha,beta)	Beta分佈中x所對應的密度 Beta分佈中q所對應的累積機率 Beta分佈中機率值為p所對應的分位點 利用Beta分佈產生n個隨機亂數
$\boldsymbol{t}$分佈 $\mathcal{T}_{\mathrm{df}}$	dt(x,df) pt(q,df) qt(p,df) rt(n,df)	$\boldsymbol{t}$分佈中x所對應的密度 $\boldsymbol{t}$分佈中q所對應的累積機率 $\boldsymbol{t}$分佈中機率值為p所對應的分位點 利用$\boldsymbol{t}$分佈產生n個隨機亂數
$\mathcal{F}$分佈 $\mathcal{F}_{(\mathrm{df1,df2})}$	df(x,df1,df2) pf(q,df1,df2) qf(p,df1,df2) rf(n,df1,df2)	$\mathcal{F}$分佈中x所對應的密度 $\mathcal{F}$分佈中q所對應的累積機率 $\mathcal{F}$分佈中機率值為p所對應的分位點 利用$\mathcal{F}$分佈產生n個隨機亂數

*指令中的“ncp”代表非中心參數(non-centrality parameter), 預設值為0。

表5.2 其他分佈中x所對應的密度之指令表

分佈名稱	指令
均勻分佈	dunif(x,min,max)
超幾何分佈	dhyper(x,m,n,k)
負二項分佈	dnbinom(x,size,prob,mu)
柯西分佈	dcauchy(x,location,scale)
指數分佈	dexp(x,rate)
對數常態分佈	dlnorm(x,meanlog,sdlog)
韋伯分佈	dweibull(x,shape,scale)
Logistic分佈	dlogis(x,location,scale)
SignRank分佈	dsignrank(x,n)
Wilcoxon分佈	dwilcox(x,m,n)

*如同表5.1, 表5.2也可有累積機率(p), 分位數(q)以及隨機亂數(r)的語法, 只須將d取代為p, q及r即可。

5.12 使用統計軟體R

我們希望將不同分佈的p.d.f.變化情形繪製在同一圖形上, 讓讀者能很清楚的看出圖形的改變, 故利用"function", "outer"以及"matplot"來達到此效果。

底下我們僅以二項分佈p.d.f.的指令作介紹, 其餘分佈只列出各分佈圖形產生的指令, 詳細的用法讀者可參考二項分佈的說明。

5.2節 繪製二項分佈p.d.f.的圖形

1. 產生1到50的數列。

   > x<−0:50

2. 利用"dbinom"計算出二項分佈p.d.f.的值, 將資料存成矩陣的型式之後, 對矩陣的行資料畫出其圖形。

   > matplot(x, outer(x, c(0.1,0.5,0.9), function(x,p) dbinom(x,50,p)),

   + lty = 1:3, type = c("h","h", "h"),ylab="")

   (a) function：自訂函數。用法為"function(自訂變數名稱) 表示式或公式"。

   (b) outer：利用函數(function)產生一個矩陣, 亦即產生$(\boldsymbol{x}, \boldsymbol{y})$。用法為"outer(x,y, function)"。

   (c) matplot：將矩陣$(\boldsymbol{x},\boldsymbol{y})$依據行的資料, 將其畫成圖形。用法為"matplot($\boldsymbol{x}$, $\boldsymbol{y}$, 其餘圖形的設定指令)"。在此$\boldsymbol{y}$是利用"outer"指令來產生。

3. 圖例說明。

   >legend(20,0.19,c("B(50,0.1)", "B(50,0.5)","B(50,0.9)"), lty=1:3,

   + col=1:3)

   利用"legend(x軸位置,y軸位置,圖例名稱,其餘圖形的指令)"指令標示出線段所各自代表的分佈名稱。

5.3節 繪製幾何分佈p.d.f.的圖形

```
> x<-0:50
> matplot(x, outer(x, c(0.1,0.5), function(x,p) dgeom(x,p)),col=1,
+ lty = 1:2, type = c("p","p"),ylab="",pch=c(1,20))
> legend(20,0.5,c("Ge(0.1)","Ge(0.5)"),col=1,pch=c(1,20))
```

5.4節 繪製波松分佈p.d.f.的圖形 繪製波松分佈的機率密度函數圖形

```
> x<-0:10
> matplot(x, outer(x, c(0.5,1,5), function(x,p) dpois(x,p)),
+ lty = 1:3, type = c("h","h","h"),ylab="",col=1)
> legend(8,0.6,c("P(0.5)","P(1)","P(5)"),lty=1:3,col=1)
```

5.5節 繪製Gamma分佈p.d.f.的圖形

```
> x<-seq(0,10,0.01)
> matplot(x, outer(x, c(0.5,1,1.5,2), function(x,alpha) dgamma(x,alpha,
+ 1/2)),col=1,lty = 1:4, type = c("l","l","l","l"),ylim=c(0,1))
> legend(5,1,c("Gamma(0.5,1/2)","Gamma(1,1/2)","Gamma(1.5,1/2)",
+ "Gamma(2,1/2)"),lty=1:4,col=1)
```

5.7節 繪製常態分佈p.d.f.的圖形

1.改變參數“σ”。

```
> x<-seq(-4,4,0.01)
> matplot(x, outer(x, c(0.5,1,2), function(x,sd) dnorm(x,0,sd)),
+ lty = 1:3, type = c("l","l", "l"),ylab="",col=1)
> legend(1.5,0.8,c("N(0,0.5)", "N(0,1)","N(0,2)"), lty=1:3,col=1)
```

2.改變參數“μ”。

```
> matplot(x, outer(x, c(-1,0,2), function(x,mu) dnorm(x,mu,1)),col=1,
+ lty = 1:3, type = c("l","l","l"),ylab="")
```

```
> legend(-4,0.4,c("N(-1,1)","N(0,1)","N(2,1)"),lty=1:3,col=1)
> legend(-4,0.4,c("N(-1,1)", "N(0,1)","N(2,1)"), lty=1:3,col=1)
```

5.8節 繪製Beta分佈p.d.f.的圖形

```
> x<-seq(0,1,0.01)
> matplot(x, outer(x, c(1,2,0.5), function(x,beta) dbeta(x,1,beta)),col=1,
+ lty = 1:3, type = c("l","l", "l"),ylab="",ylim=c(0,5))
> matlines(x, outer(x,c(3), function(x,alpha) dbeta(x,alpha,2)),col=2,
+ lty =4, type ="l",ylab="",ylim=c(0,5))
```

利用“matlines”指令增加不同分佈的圖形。

```
> matlines(x, outer(x,c(4), function(x,alpha) dbeta(x,alpha,4)),col=3,
+ lty =5, type ="l",ylab="",ylim=c(0,5))
> matlines(x, outer(x,c(1/2), function(x,alpha) dbeta(x,alpha,1/3)),col=4,
+ lty =6, type ="l",ylab="",ylim=c(0,5))
> legend(0,5,c("Beta(1,1)","Beta(1,2)","Beta(1,1/2)","Beta(3,2)",
+ "Beta(4,4)","Beta(1/2,1/4)"), lty=1:6,col=c(1,1,1,2,3,4))
```

5.9節 繪製T分佈p.d.f.的圖形

```
> x<-seq(-4,4,0.01)
> matplot(x, outer(x, c(1,2,10,30), function(x,n) dt(x,n)),col=1,
+ lty = 1:4, type = c("l","l", "l","l"), ylab=" ")
> legend(2,0.4,expression(plain(T)[1],plain(T)[2],plain(T)[10],
+ plain(T)[30]),lty=1:4,col=1)
```

利用“expression”指令可達到數學上常用的表示法, 如π或下標。

5.10節 繪製F分佈p.d.f.的圖形

```
> x<-seq(0,5,0.01)
> matplot(x, outer(x, c(1,2,5,10,20), function(x,n) df(x,10,n)),col=1,
+ lty = 1:5, type = c("l","l", "l","l", "l"),ylab="")
```

```
> legend(3,0.8,c("F(10,1)", "F(10,2)","F(10,5)", "F(10,10)","F(10,20)"),
+ lty=1:5,col=1)
```

習　題

1. 利用指數分佈的指令,分別給出參數為1,2,3之指數分佈的圖形。
2. 標準常態分佈之分佈函數為$\Phi(z_p) = P(Z \leq z_p) = p$,分別給出$p = 0.9, 0.95, 0.975, 0.99$之值。

第六章

假設檢定

6.1 前言

在本章中, 將透過統計上常用的檢定方法, 來了解數據中所隱含的訊息, 並對結果作一簡單的分析。

6.2 單樣本之t檢定

對一組樣本, 想知道是否取自一個平均值為$\mu = \mu_0$的常態分佈, 可採用單樣本之$\boldsymbol{t}$檢定。此時虛無假設取為$H_0 : \mu = \mu_0$。

設$X_1, X_2, \cdots, X_n$為由$\mathcal{N}(\mu, \sigma^2)$分佈所產生之隨機樣本, 且設$\sigma^2$為已知。則

$$Z_0 = \frac{\overline{\boldsymbol{X}} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{\boldsymbol{X}} - \mu_0)}{\sigma}。$$

在虛無假設H_0下, Z_0有$\mathcal{N}(0,1)$分佈。如果母體並非常態分佈, 但樣本數n夠大, 則Z_0有近似的常態分佈。

若σ^2為未知, 可以樣本標準差s來替代σ。則在虛無假設H_0下,

$$t = \frac{(\overline{\boldsymbol{X}} - \mu)}{s/\sqrt{n}} = \frac{\sqrt{n}(\overline{\boldsymbol{X}} - \mu)}{s},$$

有T_{n-1}分佈。

我們利用常態分佈生成隨機亂數並進行檢定分析, 分別為$\mathcal{N}(0,1)$, $\mathcal{N}(5,1)$, $\mathcal{N}(5,5)$, 每組資料各有10筆數據。資料內容如表6.1。

表6.1 由常態分佈生成之隨機亂數數據

	$\mathcal{N}(0,1)$	$\mathcal{N}(5,1)$	$\mathcal{N}(5,5)$
1	1.00405702	5.82325516	2.85546774
2	-0.37944062	3.65857746	9.41795098
3	-1.33690441	5.08153128	-2.51473331
4	1.70595851	4.63310514	2.78444268
5	-1.021613848	4.78817883	-4.51136987
6	0.02488106	5.578173303	0.15908764
7	-2.03636920	5.74009755	6.02712745
8	0.27912606	4.60472340	8.43356766
9	-0.04036002	4.94552555	12.54101082
10	1.42386970	4.79023810	3.17784588

先繪製表6.1之資料的盒鬚圖及常態分位圖, 如圖6.1及6.2。首先利用表6.1的$\mathcal{N}(0,1)$ 數據來進行單樣本之t檢定分析。

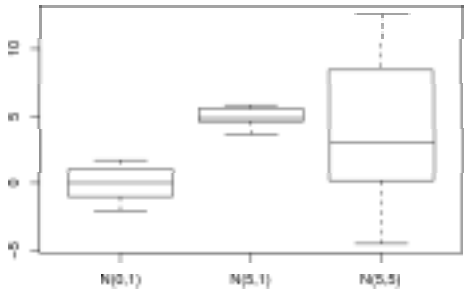


圖 6.1 由常態分佈生成之隨機亂數數據的盒鬚圖

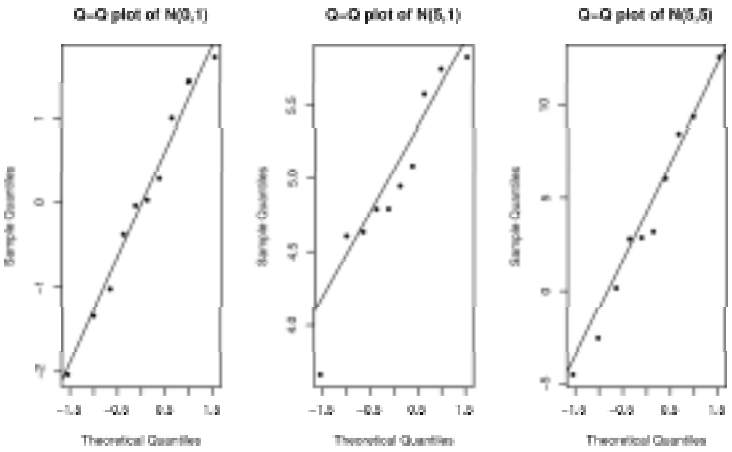
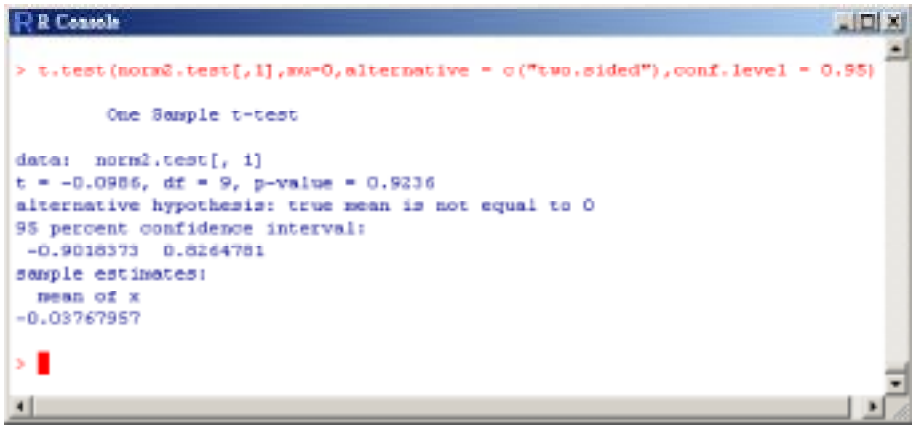


圖 6.2 由常態分佈生成之隨機亂數數據的常態分位圖

利用指令“t.test”, 可得圖6.3的分析結果。若一開始並不知道此數據由$\mathcal{N}(0,1)$分佈所產生的, 由單樣本之$\boldsymbol{t}$檢定的結果可知其p值(p-value)為0.9236, 此p值大於所設定的顯著水準0.05。故由檢定結果知接受虛無假設$H_0: \mu = 0$。在分析結果中, 也列出了95%信賴區間及樣本平均。



```
R Console
> t.test(norm2.test[,1],mu=0,alternative = c("two.sided"),conf.level = 0.95)

        One Sample t-test

data:  norm2.test[, 1]
t = -0.0986, df = 9, p-value = 0.9236
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.9018373  0.8264781
sample estimates:
  mean of x
-0.03767957

>
```

圖 6.3 單樣本$\boldsymbol{t}$檢定

6.3 二樣本之t檢定

在很多實際的應用中, 常會遇到二筆數據樣本平均之差的檢定問題。譬如說, 有某一特別的教學法, 想知道上過此教學法的學生之平均成績, 是否會較未上過該種教學法的平均成績高？二樣本之t檢定即是用於檢定配對中二樣本之平均值是否相同, 或兩個具有相同標準差之常態分佈是否有相同的平均值。

假設有兩組獨立的隨機樣本$X_1,\cdots,X_{n_1}$及$Y_1,\cdots,Y_{n_2}$, 分別有$\mathcal{N}(\mu_1,\sigma_1^2)$及$\mathcal{N}(\mu_2,\sigma_2^2)$分佈。令$\overline{X}$, S_1^2及$\overline{Y}$, S_2^2分別表第一組及第二組樣本之樣本平均及樣本變異數。則

$$Z=\frac{(\overline{Y}-\overline{X})-(\mu_2-\mu_1)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \tag{6.1}$$

有$\mathcal{N}(0,1)$分佈。又

$$V=\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2_{n_1+n_2-2}。$$

則

$$T_1=\frac{Z}{\sqrt{V/(n_1+n_2-2)}}\sim\mathcal{T}_{n_1+n_2-2}。$$

若σ_1^2,σ_2^2未知且不相等, 則T並非統計量。但若$\sigma_1^2=\sigma_2^2$, 則$T_1=T$, 其中

$$T=\frac{(\overline{Y}-\overline{X})-(\mu_2-\mu_1)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{6.2}$$

為一有$\mathcal{T}_{n_1+n_2-2}$分佈之統計量。利用T來做檢定稱為二樣本之t檢定(two-sample t test)。

由表6.1的$\mathcal{N}(0,1)$及$\mathcal{N}(5,1)$之數據來進行二樣本之t檢定分析。指令與單樣本之t檢定為同一指令, 即“t.test”。若一開始並不知道此二筆數據, 是由$\mathcal{N}(0,1)$及$\mathcal{N}(5,1)$分佈所產生的, 且假設變異數是相等的, 利用t檢定來作分析, 可得圖6.4的分析結果。其p值為9.268e-10, 此p值小於所設定的顯著水準0.05。故由檢定結果拒絕虛無假設$H_0:\mu_1=\mu_2$。同樣的, 在分析結果

中，也列出了95%信賴區間及二組資料的樣本平均。

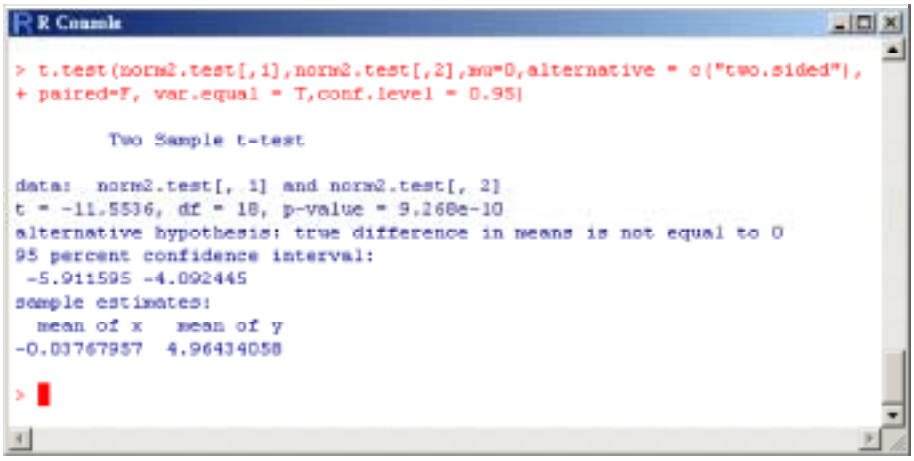


圖 6.4 二樣本之t檢定，假設變異數是相等的

若假設變異數是不相等的，可得圖6.5的分析結果，其p值為1.856e-8，此p值小於所設定的顯著水準0.05。故由分析結果知此數據同樣拒絕虛無假設$H_0 : \mu_1 = \mu_2$。同時也在分析結果中，列出了95%信賴區間及二組資料的樣本平均。

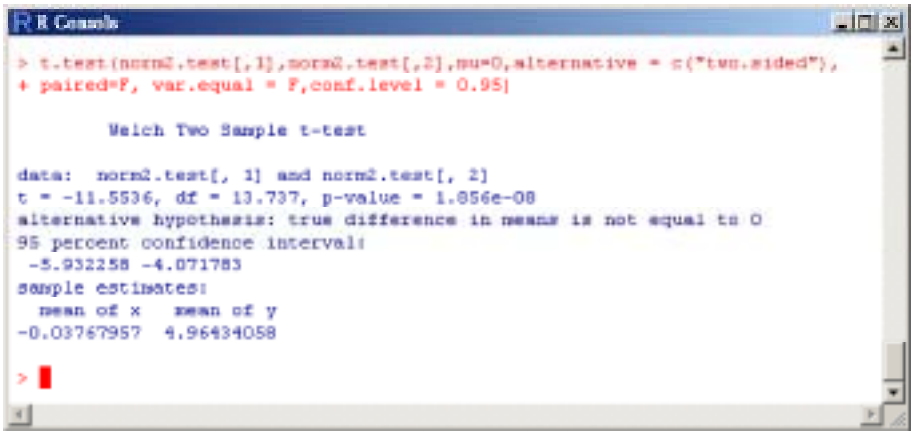


圖 6.5 二樣本之t檢定，假設變異數是不相等的

6.4 變異數的檢定

母體變異數之檢定也是很重要的, 期望值如果是目標值, 則變異數量測數據偏離目標值之情況。所以不論是有關產品之品質, 或考題的製作, 都會很在乎變異數的大小。對於產品的品質, 通常會希望變異數不要超過某一定值。至於對一特定的測驗, 為了有鑑別效果, 有時會預設變異數, 或希望變異數超過某一定值。另外, 兩種不同的方法所產生的變異數也是常需要比較的。F-檢定亦稱為變異數比例檢定。F-檢定是一種用於二組樣本之變異數的檢定。

設有兩組獨立的隨機樣本$X_1,\cdots,X_{n_1}$, $Y_1,\cdots,Y_{n_2}$, 分別有$\mathcal{N}(\mu_1,\sigma_1^2)$及$\mathcal{N}(\mu_2,\sigma_2^2)$ 分佈。又以S_1^2, S_2^2分別表兩組樣本之樣本變異數。欲檢定$H_0:\sigma_1^2/\sigma_2^2=1$。

先設μ_1,μ_2皆為已知。則

$$\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/(n_1\sigma_1^2)}{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2/(n_2\sigma_2^2)}\sim\mathcal{F}_{n_1,n_2},$$

若μ_1,μ_2皆為未知, 而H_0為眞時, 則

$$\frac{S_1^2}{S_2^2}\sim\mathcal{F}_{n_1-1,n_2-1}。$$

利用表6.1的$\mathcal{N}(0,1)$及$\mathcal{N}(5,5)$生成之數據來進行變異數的檢定分析。指令為“var.test”。若一開始並不知道此二筆數據, 是由$\mathcal{N}(0,1)$及$\mathcal{N}(5,5)$分佈所產生的, 利用變異數檢定分析可得圖6.6的分析結果。其p值為0.0001373, 此p值小於所設定的顯著水準0.05。由分析結果知此數據拒絕虛無假設$H_0:\sigma_1^2/\sigma_2^2=1$。同樣的, 在分析結果中, 亦列出了95%信賴區間及二組資料之樣本變異數的比例值。

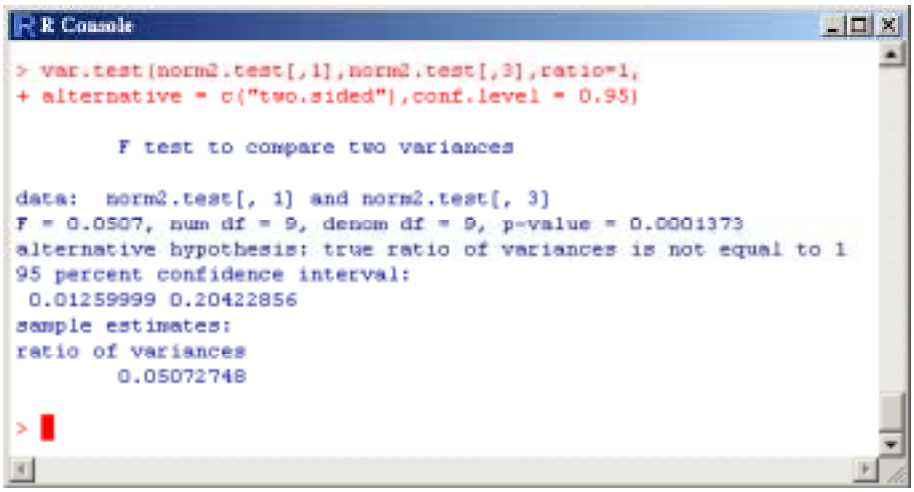


圖 6.6 變異數的檢定

6.5 卡方檢定

有時在作分析之前, 對所收集到的數據會有一主觀的猜測, 像是前幾節的例子, 是由常態分佈所生成的隨機亂數資料, 所以我們知道應該會符合某些性質。但是對於收集到的數據, 分析者無法獲得這麼完整的訊息, 可以知道源自何種分佈。一般皆是根據過去經驗, 或由已知母體的一些特性, 猜測所收集到的資料應該符合某一分佈, 可是這樣的主觀猜測有時必須先檢定其是否正確。在這一節中, 要介紹一有用且簡易的方法。

對一隨機現象, 由收集到的數據, 想檢定某一模型對此數據是否合適, 或想區別數據的偏差是由於機運所產生, 或真的是因模型不夠精確, 這種過程便統稱為適合度檢定(goodness of fit)。卡方檢定(χ^2 test)又稱為皮爾生之卡方適合度檢定 (Pearson's chi-squared goodness-of-fit test)是最常用的適合度檢定法之一。

皮爾生卡方檢定統計量為

$$Q = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

其中觀測頻率以o_i表之, 期望頻率以e_i表之。

以孟德爾著名的關於豌豆生長的實驗為例。他將有圓黃(round yellow)種子的與有縐綠(wrinkled green)種子的豌豆雜交。依其理論, 會生長出圓黃、圓綠、縐黃及縐綠種子的後代之比率, 應分別為9/16, 3/16, 3/16及1/16。經由一組有556個樣本的實驗, 我們列出觀測頻率及期望頻率於表6.2。

表6.2 豌豆雜交後生長的觀測頻率及期望頻率

	圓黃	圓綠	縐黃	縐綠
觀測頻率	315	108	101	32
期望頻率	312.75	104.25	104.25	34.75

由表6.2的資料, 利用R的卡方檢定之指令“chisq.test”, 所得結果$Q = 0.47$, p值為0.9254, 參考圖6.7。因此無法拒絕虛無假設H_0:孟德爾的理論為正確。

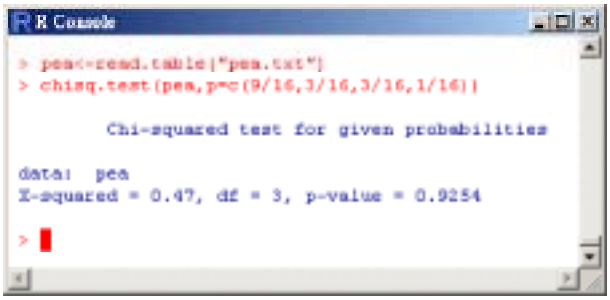


圖 6.7 孟德爾之豌豆生長實驗的卡方檢定

6.6 無母數檢定

在這一節中, 討論當我們不知道所收集的資料來自何種分佈時, 如何針對資料做出一些推論。我們引進無母數檢定的方法, 來解決此問題。

無母數檢定的方法有相當多種, 有興趣的讀者可以參考一般的無母數統計的書籍, 如Conover(1999)。底下我們介紹其中的一種檢定, 稱為單樣本之Kolmogorov-Smirnov檢定, 亦可稱為單樣本之K-S檢定。

Kolmogorov-Smirnov檢定通常是用來檢定一隨機樣本是否來自某一特定分佈的適合度檢定。假設有一組來自分佈函數為$F(x)$之隨機樣本$X_1, \cdots, X_n$。想要知道此n個樣本是否來自一分佈函數為$F^*(x)$之分佈。一個常用來與$F^*(x)$比較的值為$S(x)$, $S(x)$稱為$X_1, \cdots, X_n$之經驗分佈函數(empirical distribution function),定義為

$$S(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \le x\}}, \quad -\infty < x < \infty。$$

由$S(x)$的定義, 可知$E(S(x)) = F(x)$, $Var(S(x)) = \frac{F(x)(1-F(x))}{n}$。我們想知道$S(x)$與$F^*(x)$兩者是否一致。一個簡單的方式就是比較兩者差的最大值,即令$T(x) = \sup_x |F^*(x) - S(x)|$,並以$T(x)$為檢定統計量。對檢定

$$H_0 : F(x) = F^*(x), \quad \forall x \in (-\infty, \infty),$$
$$H_a : F(x) \neq F^*(x), \quad 至少存在一x。$$

表6.3給出一些Kolmogorov檢定統計量之對應的分位數值表。如果$T(x)$大於$1-\alpha$分位數, 就拒絕H_0的假設。

我們可利用"ks.test"指令求得Kolmogorov-Smirnov檢定之結果。假設我們有如下的10筆數據0.621, 0.503, 0.203, 0.477, 0.710, 0.581, 0.329, 0.480, 0.554, 0.382。想要知道此組數據是否來自0到1之間的均勻分佈。圖6.8為此組數據之經驗分佈函數與均勻分佈的圖形。利用"ks.test"得Kolmogorov-Smirnov檢定之結果, 如圖6.9。檢定統計量$T(x) = 0.29$, 且p值$= 0.3696$, 只要$\alpha < 0.3696$,檢定結果接受虛無假設。

表6.3 Kolmogorov檢定統計量之對應的分位數值表

n	$p=0.80$	$p=0.90$	$p=0.95$	$p=0.98$	$p=0.99$
1	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.557	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	0.409	0.457	0.489

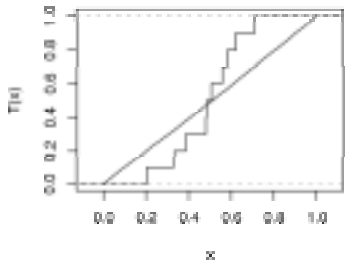


圖 6.8 經驗分佈函數與均勻分佈之圖形

圖 6.9 Kolmogorov-Smirnov檢定之結果

6.7 使用統計軟體R

6.2節 單樣本之T 檢定

1. 生成表6.1的常態分佈之隨機亂數資料。

```
> norm2.test<-data.frame(rnorm(10,mean=0,sd=1),
+ rnorm(10,mean=5,sd=1),rnorm(10,mean=5,sd=5))
```

2. 圖6.1之盒鬚圖。

```
> boxplot(norm2.test)
```

3. 圖6.2之常態分位圖

```
> par(mfrow=c(1,3))
> qqnorm(norm2.test[,1],main="Q-Q plot of N(0,1)",pch=20)
> qqline(norm2.test[,1])
> qqnorm(norm2.test[,2],main="Q-Q plot of N(5,1)",pch=20)
> qqline(norm2.test[,2])
> qqnorm(norm2.test[,3],main="Q-Q plot of N(5,5)",pch=20)
> qqline(norm2.test[,3])
```

4. 圖6.3單樣本之t檢定。

```
> t.test(norm2.test[,1],mu=0,alternative = c("two.sided"),
+ conf.level = 0.95)
```

其中"mu"是設定虛無假設$H_0 : \mu = 0$。"alternative"設定檢定採用雙邊, 上單邊或下單邊檢定。"conf.level"為設定顯著水準的值。

6.3節 二樣本之t檢定

1. 圖6.4二樣本之t檢定–假設變異數相同。

```
>t.test(norm2.test[,1],norm2.test[,3],alternative = c("two.sided"),
```

+ paired=F,var.equal = T,conf.level = 0.95)

其中“paired”是判斷資料是否為成對數據，預設值為FALSE。“var.equal”是設定此檢定之變異數是否相同，預設值為FALSE。

2. 圖6.5二樣本之t檢定–假設變異數不同。

> t.test(norm2.test[,1],norm2.test[,3],alternative = c("two.sided"),

+ paired=F,var.equal = F,conf.level = 0.95)

6.4節 變異數檢定

1. 圖6.6變異數的檢定。

> var.test(norm2.test[,1],norm2.test[,3],ratio=1,

+ alternative = c("two.sided"),conf.level = 0.95)

其中“ratio”是設定虛無假設$H_0 : \sigma_1^2/\sigma_2^2 = 1$”，亦即二變異數之比例值是否為1之檢定。

6.5節 卡方檢定

1. 圖6.7卡方檢定。

> pea<−read.table("pea.txt")

> chisq.test(pea,p=c(9/16,3/16,3/16,1/16))

其中chisq.test指令中的“p”代表樣本出現的機率，非期望頻率。

6.6節 無母數檢定

1. 輸入數據資料。

> t<−c(0.621,0.503,0.203,0.477,0.710,0.581,0.329,0.48,0.554,0.382)

2. 求Kolmogorov-Smirnov檢定。

> ks.test(t, "punif ", 0, 1)

> plot.stepfun(runif(10000,0,1), col.vert = "gray20",do.points =F,

```
+ main="",ylab="T(x)")
> plot.ecdf(t,verticals =T,xlim=c(0,1),add=T,do.points =F)
```

指令中之“plot.stepfun”代表繪製出階梯函數之圖形, “col.vert ”設定垂直線之顏色, “do.points”是否在連接處標上圓點, “plot.exdf”繪出經驗分佈函數之圖形, “verticals”是否畫出垂直的連接線。

習　題

1. 底下資料分別代表機器和人工所量得之血壓值。分別討論其T及F檢定的結果。

機器量測值

68	82	94	106	92	80	76	74
110	93	86	65	74	84	100	79

人工量測值

72	84	89	100	97	88	84	70
103	84	86	63	69	87	93	84

2. 底下資料為台灣銀行發行的共24期公益彩券幸運獎中獎號碼之末6位數碼$0, 1, \cdots, 9$的觀測頻率。欲檢定其是否均勻出現。

數字	0	1	2	3	4	5	6	7	8	9
觀測頻率	12	13	13	21	8	16	16	18	13	14

3. 由底下兩筆數據, 判斷此兩筆數據是否來自同一分佈。

第一筆數據

0.165	0.063	-0.616	0.748	-1.457	-0.288	2.104	0.082
1.531	-1.236	0.196	-0.228	0.544	-0.405	0.171	1.208
-0.503	0.394	-0.791	0.035	-1.236	0.098	1.5	-1.003
-1.572	1.749	-1.29	0.766	-0.475	-1.191		

第二筆數據

0.542	0.893	0.614	0.358	0.674	0.030	0.870	0.759
0.362	0.716						

第七章

估計

7.1 最大概似估計法

假設已知樣本$\boldsymbol{x}$源自於機率密度函數為$f(x|\theta)$之分佈，但其中之參數θ未知。θ的值會影響到之後所作的任何分析或推論。因此如何估計θ便成為一很重要的課題。

最大概似估計法(method of maximum likelihood)，就是找參數θ之估計量$\hat{\theta}$，使得在$\theta=\hat{\theta}$之下，會最可能產生數據$\boldsymbol{x}$。本來要先給出θ，樣本$\boldsymbol{x}$的分佈才會完全決定。現由觀測到的值$\boldsymbol{x}$倒回去推想，怎樣的θ，才會使此$\boldsymbol{x}$在諸多可能的$\boldsymbol{x}$中拔得頭籌？若$\hat{\theta}$存在，便稱為θ之最大概似估計值(maximum likelihood estimate，簡寫為MLE)。最大概似估計的理論架構及性質，請參考一般數理統計書籍，如黃文璋(2003)。

想要利用R幫忙找出最大概似估計值，須先下載資料庫裡的“MASS”，參考圖7.1及圖7.2。

利用最大概似估計法可推得一些常見分佈的最大概似估計值。以$\mathcal{N}(\mu,\sigma^2)$來說，設有$n$個樣本，$x_1,\cdots,x_n$。若$\mu$及$\sigma^2$皆未知，則可推得$\mu$及$\sigma^2$的最大概似估計值分別為$\hat{\mu}=\bar{x}_n$及$\hat{\sigma}^2=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2/n$。而參數為$\lambda$的指數分佈之最大概似估計值則為$\hat{\lambda}=1/\bar{x}_n$。我們以此二分佈為例，利用R的

指令“fitdistr”分別求出此二分佈的最大概似估計值。

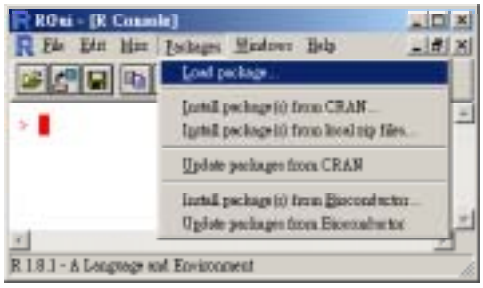


圖 7.1 開啓“下載資料庫”的畫面

圖 7.2 選擇欲下載資料庫檔案的畫面

利用R生成10筆$\mathcal{N}(0,1)$的數據, 如表7.1。由R的指令計算後可得此筆資料的$\hat{\mu} = -0.008086072$及$\hat{\sigma}^2 = 0.839037210$, 由公式推導的結果$\mu$及$\sigma^2$之最大概似估計值分別為$\hat{\mu} = \bar{x}_n = -0.008086072$及$\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2/n = 0.839037210$, 估計結果及公式驗算參考圖7.3。

同樣地, 利用R生成10筆指數分佈參數$\lambda = 0.5$的數據, 如表7.2。透過R的計算可得$\hat{\lambda} = 0.6530994$, 由公式推導的結果$\lambda$之最大概似估計值為$\hat{\lambda} = 1/\bar{x}_n = 0.6531623$, 估計結果及公式驗算如圖7.4。

從求常態分佈及指數分佈參數的最大概似估計值的結果中, 可看出此指令為數值計算的結果會與實際利用公式推導出的估計值有些微誤差。在常態分佈的例子中, 不論公式計算或透過R來求最大概似估計值皆可得到

同樣的結果, 但是在指數分佈的例子中, 兩個估計值有些微的誤差, 兩者的差為0.000629。當樣本數夠多時此誤差值會越小。利用指令“fitdistr”可求得常態分佈、指數分佈、Gamma分佈、負二項分佈、Beta分佈、對數常態分佈、韋伯分佈等的最大概似估計值。

表7.1 由$\mathcal{N}(0,1)$生成之10筆數據

1	-1.25494872994457	2	0.580494591832429
3	-0.203463056389231	4	0.0507923242732363
5	-0.400148582512666	6	-0.133561998301612
7	-1.48978604201467	8	0.883319416939335
9	1.19584395609907	10	0.690597401520781

表7.2 由參數$\lambda = 0.5$的指數分佈生成之10筆數據

1	6.18360018696556	2	0.0952219185649138
3	0.92688906379044	4	0.783767773644048
5	0.849087740046285	6	0.37164671998471
7	2.23702448970618	8	0.509015057058229
9	0.381319892592728	10	2.97255648867447

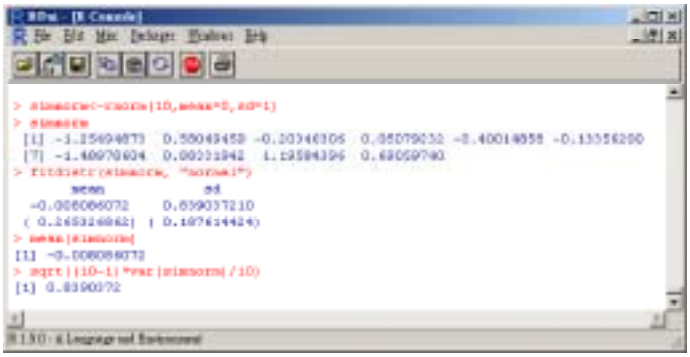


圖 7.3 $\mathcal{N}(0,1)$最大概似估計的結果與公式驗算

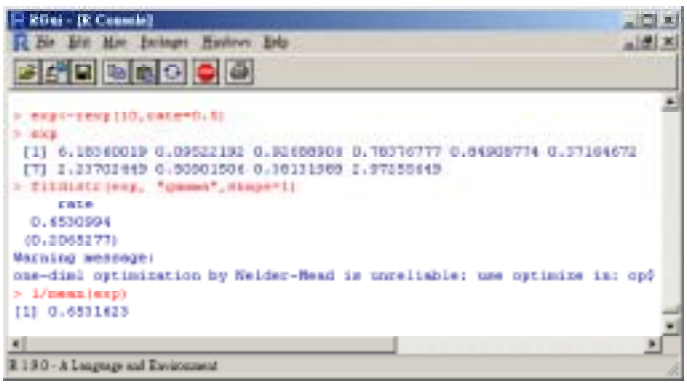


圖 7.4 參數$\lambda = 0.5$之指數分佈最大概似估計的結果與公式驗算

7.2 區間估計

對於一參數, 除了最大概似估計值, 有時我們想對估計值與眞實值究竟多接近能有些了解。在這一節中我們給出常態分佈期望值的區間估計(interval estimation), 至於推導過程及其他區間估計的討論, 請參考一般數理統計書籍, 如黃文璋(2003)。

設$X_1, \cdots, X_n$為一組由$\mathcal{N}(\mu, \sigma^2)$分佈所產生之隨機樣本, 欲給出$\mu$之信賴區間。先設$\sigma^2$已知。由於樣本平均$\overline{X}$有$\mathcal{N}(\mu, \sigma^2/n)$分佈, 因此$(\overline{X} - \mu)/(\sigma/\sqrt{n})$有$\mathcal{N}(0,1)$分佈, 故

$$P_\mu\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right) = 1 - \alpha, \tag{7.1}$$

與μ無關。即得

$$\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{7.2}$$

為μ之一信賴係數為$1-\alpha$之信賴區間。若σ^2未知, 則以

$$\left[\overline{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] \tag{7.3}$$

為μ之$1-\alpha$信賴區間, 其中$S = \sqrt{S^2} = \sqrt{\sum_{i=1}^n (X_i - \overline{X})^2/(n-1)}$。

我們以表7.1的數據為例，在此R的指令與第六章t檢定的指令相似。假設此筆數據的μ及σ^2皆未知，在第六章中我們利用“t.test”指令可給出單樣本之t檢定的結果，在此我們只須多加一些指令就可將此筆數據的信賴區間求出，此指令即為“t.test$conf.int”。圖7.5給出此筆數據的信賴區間及公式驗算結果。

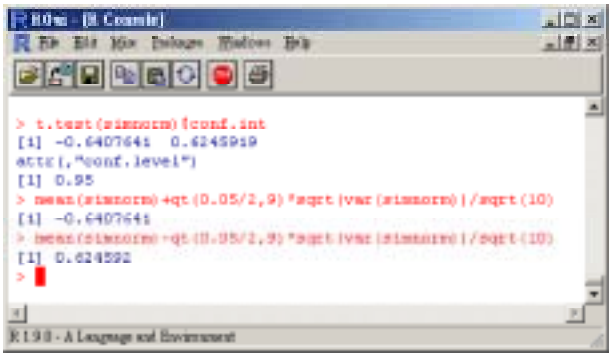


圖 7.5 常態分佈期望值的信賴區間結果

順帶一提，利用同樣的方式也可只列出“t.test”檢定結果的單一個值，只須將“conf.int”改成所要求之值的指令即可，對應的指令參考表7.3。

表7.3 有關“t.test”的一些細部指令

指令	對應指令的意義
statistic	檢定統計量的值
parameter	t檢定的自由度
p.value	檢定的p值
conf.int	信賴區間
estimate	參數的估計值
null.value	虛無假設的檢定內容，檢定平均值或二平均值的差
alternative	虛無假設採用單邊或雙邊檢定
method	檢定方法，單樣本檢定或雙樣本檢定
data.name	檢定的資料名稱

7.3 使用統計軟體R

7.1節 最大概似估計法

1. 表7.1生成10筆$\mathcal{N}(0,1)$的數據。

   > simnorm<−rnorm(10,mean=0,sd=1)

2. 圖7.3求出μ及σ的最大概似估計值。

   > fitdistr(simnorm, "normal")

3. 直接代入常態分佈最大概似估計值的公式。

   > mean(simnorm)

   > (10-1)∗ sd(simnorm)/10

4. 表7.2生成10筆參數λ =0.5的指數分佈的數據。

   > exp<−rexp(10,rate=0.5)

5. 圖7.4利用Gamma(1,1/λ)模型求出λ的最大概似估計值。

   > fitdistr(exp, "gamma",shape=1)

   在此利用到指數分佈為一特別的Gamma分佈, 所以在下指令時透過Gamma分佈的最大概似估計值求法, 可得指數分佈的最大概似估計值。

6. 直接代入指數分佈最大概似估計值的公式。

   > 1/mean(exp)

7.2節 區間估計

1. 圖7.5利用T檢定的指令求出信賴區間。

   > t.test(simnorm)$conf.int

2. 直接代入公式求出區間值。

```
> mean(simnorm)+qt(0.05/2,9)* sqrt(var(simnorm))/sqrt(10)
> mean(simnorm)-qt(0.05/2,9)* sqrt(var(simnorm))/sqrt(10)
```

習　題

1. 利用分佈的相關指令,生成一組樣本數為20之標準常態分佈的數據,並給出其最大概似估計值。
2. 利用上題所生成之數據,給出數據的信賴區間及p值。

第八章

迴歸分析

8.1 簡單線性迴歸

在簡單線性迴歸中, 我們有下述關係式

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{8.1}$$

其中Y為隨機變數, x為一獨立的變數。而β_0, β_1為二待估的參數, 分別表迴歸線之斜率及截距。此外ε為一隨機變數, 通常假設$E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$。因此可得

$$E(Y) = \beta_0 + \beta_1 x。 \tag{8.2}$$

一般稱(8.1)式為簡單線性迴歸模型。在簡單線性迴歸問題中, 當獲得n對的觀測值, $(x_1, y_1), \cdots, (x_n, y_n)$, 則由(8.1)式可寫出樣本迴歸模型如下：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n。$$

利用"最小平方法", 可推得β_0及β_1 之估計值, 分別以$\hat{\beta}_0$及$\hat{\beta}_1$表之。在簡單線性迴歸模型中, 最小平方法是基於下述二次式

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2。 \tag{8.3}$$

(8.3)式分別對β_0及β_1微分並令其為0, 則可得(8.4)式及(8.5)式。

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0, \tag{8.4}$$

$$\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0。 \tag{8.5}$$

根據(8.4)式及(8.5)式, 解出

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}。$$

其中$\bar{x} = \sum_{i=1}^{n} x_i/n$, $\bar{y} = \sum_{i=1}^{n} y_i/n$。(詳細推導過程請參考一般線性迴歸書籍, 如Montgomery(2001))

底下我們給一例子, 利用R的指令“lsfit”, 可得其線性迴歸模型的估計值及其他相關統計值, 如p值等。McDonald and Studden(1990)曾研究汽車排出碳氫化合物(hydrocarbon) 的量。他們得到下述成對總哩程數x(千英哩)與碳氫化合物y(公克/英哩)的值如下:

$$(5.133, 0.265), (10.124, 0.287), (15.060, 0.282), (19.946, 0.286),$$
$$(24.899, 0.310), (29.792, 0.333), (29.877, 0.343), (35.011, 0.335),$$
$$(39.878, 0.311), (44.862, 0.345), (49.795, 0.319)。$$

由R的指令可得此例之迴歸分析資料, 如圖8.1, 及線性迴歸函數圖形, 如圖8.2。

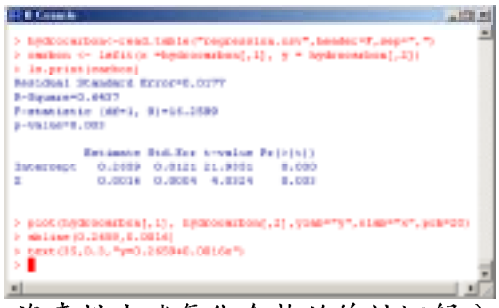


圖 8.1 汽車排出碳氫化合物的線性迴歸分析資料

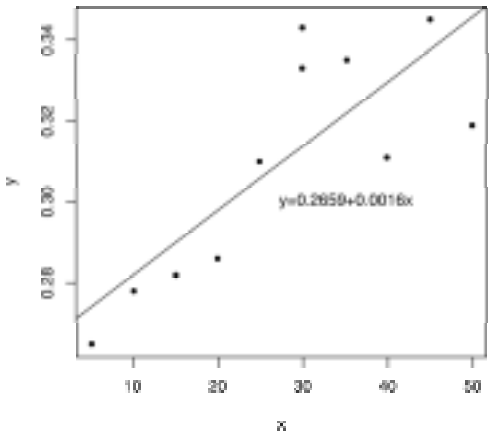


圖 8.2 汽車排出碳氫化合物的線性迴歸分析函數圖形

在前面所提之汽車排出碳氫化合物之數據分析, 僅給出所配適的簡單線性迴歸模型, 和其配適的線性迴歸函數圖形。一般的迴歸分析步驟和方法, 在前例中並沒有提及。底下以老忠實噴泉的數據為例, 作一詳細的迴歸分析。另外前面所用到的指令為"lsfit", 此指令所適用的模型為$Y = \beta X + \varepsilon$。除了lsfit指令外，亦可用"lm"來作迴歸分析, 此為配適一般線性模型的指令, 透過"lm"指令可以作出一完整的迴歸分析。完整的迴歸分析一般應包含資料的散佈圖、參數的估計值、配適結果的函數及圖形、完整的迴歸分析表、參數估計的檢定、變異數分析表、參數的信賴區間、殘差與配適值的散佈圖以及殘差的常態Q-Q圖。底下我們搭配老忠實噴泉的數據資料進行分析並給出相關資料的公式和指令。

首先, 畫出散佈圖, 我們在第三章的圖3.8已給出老忠實噴泉數據的散佈圖。由圖可看出資料似乎呈現直線的散佈狀態, 利用"lm"指令可得如圖8.3的分析結果。

由圖8.3中的分析資料可寫出配適的線性迴歸模型為

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -1.874016 + 0.075628x。$$

透過圖8.3的分析結果, 除了可得到參數的估計值外, 還有一些相關的數據，如R^2、F統計量等, 稍後將會一一介紹。接下來由估計的結果可畫出其線性迴歸模型的圖，如圖8.4。同時利用“aov”指令可得變異數分析表(analysis of variance)，如圖8.5。在此先簡單說明變異數分析表。變異數分析為一用來檢定迴歸顯著性方法。簡單線性迴歸模型之完整變異數分析表, 如表8.1。

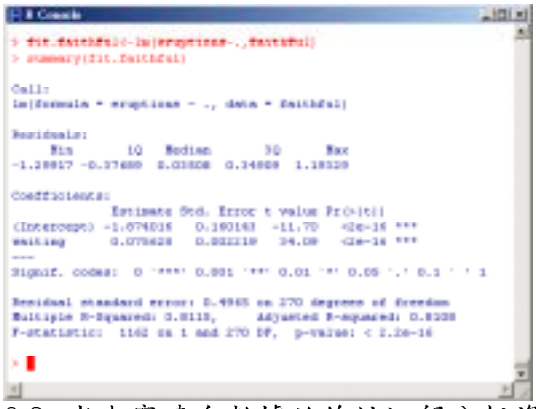


圖 8.3 老忠實噴泉數據的線性迴歸分析資料

變異數分析表中的統計量F_0可用來檢定斜率是否為0, 即檢定H_0：$\beta_1 = 0$。若要拒絕H_0, 則$F_0 > F_{\alpha,1,n-2}$。由圖8.5的資料顯示老忠實噴泉數據的$F_0 = 1162.1 > F_{0.05,1,270} = 0.003939457$, 故拒絕$H_0$的假設。

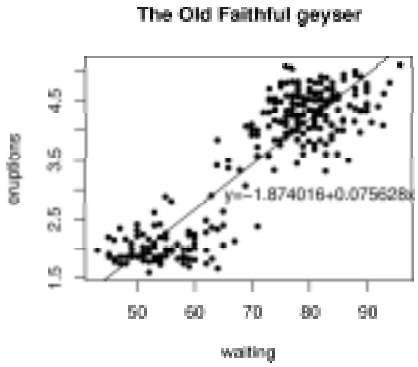


圖 8.4 老忠實噴泉數據的線性迴歸分析函數圖形

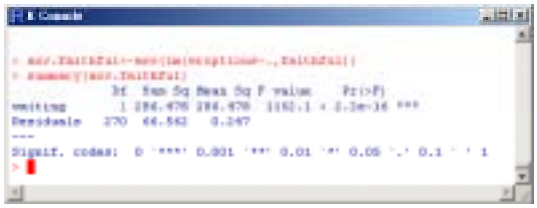


圖 8.5 老忠實噴泉數據的變異數分析結果

表8.1 變異數分析表

變異原因	平方和	自由度	均方	F_0
迴歸	$SS_R = \hat{\beta}_1 S_{xy}$	1	$MS_R = SS_R$	$\frac{MS_R}{MS_E}$
殘差	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	$MS_E = \frac{SS_E}{n-2}$	
總和	SS_T	$n-1$		

其中$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})$, $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$。

除了對β_0, β_1作估計外, 亦可給出兩者之區間估計。分別給出β_0及β_1之$100(1-\alpha)\%$信賴區間如下：

$$\hat{\beta}_0 - t_{\alpha/2,n-2}se(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}se(\hat{\beta}_0),$$

$$\hat{\beta}_1 - t_{\alpha/2,n-2}se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}se(\hat{\beta}_1),$$

其中$se(\hat{\beta}_0) = \sqrt{MS_E(1/n + \bar{x}^2/\sum_{i=1}^{n}(x_i - \bar{x})^2)}$為截距的標準差, $se(\hat{\beta}_1) = \sqrt{MS_E/\sum_{i=1}^{n}(x_i - \bar{x})^2}$為斜率的標準差。利用“confint”可得老忠實噴泉之截距與斜率之95%信賴區間, 如圖8.6。在圖8.3的分析結果中, 尚有一個統計值, 即$R^2 = SS_E/SS_T$, 稱為決定係數(coefficient of determination)表示迴歸因子x所能解釋變異的百分比, 故老忠實噴泉的$R^2 = 0.8115$, 表示透過迴歸模型能夠解釋約81.15%的變異。

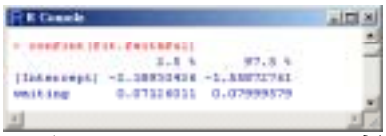


圖 8.6 老忠實噴泉截距與斜率之95%信賴區間

了解了一些相關的分析表和統計值後, 除了獲得迴歸模型外, 我們也需要知道模型配適的好不好, 此時就得透過殘差來作判斷。殘差表示觀察值y與配適值$\hat{y}$之間的差值。透過殘差的圖形分析可以知道迴歸模型是否配適恰當。圖8.7和圖8.8給出了老忠實噴泉的殘差與配適值的散佈圖和殘差的常態Q-Q圖。由常態Q-Q圖呈現一直線的情形來判斷, 此迴歸模型配適的不錯。

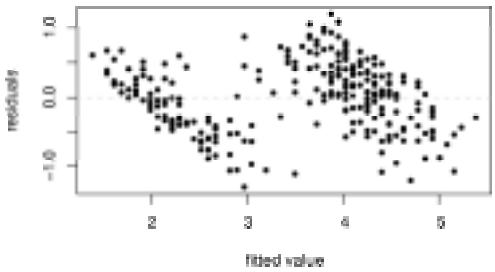


圖 8.7 老忠實噴泉數據的殘差與配適值的散佈圖

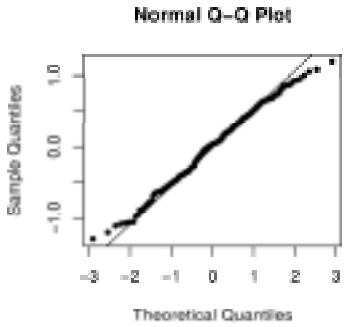


圖 8.8 老忠實噴泉數據殘差之常態Q-Q圖

8.2 單因子變異數分析

變異數分析(analysis of variance, 簡稱ANOVA)是一種常用在實驗設計資料上的分析方法。在此我們僅簡單介紹單因子變異數分析方法(one-

way ANOVA)。

單因子變異數分析的模型為

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n, \tag{8.6}$$

其中y_{ij}為第i組(treatment, 或稱因子水準(factor level))資料中第j個觀測值, μ為k組的共同參數, τ_i為影響第i組結果的一個參數, ε_{ij}為一誤差項並且滿足$E(\varepsilon_{ij}) = 0$及$Var(\varepsilon_{ij}) = \sigma^2, \forall\, i = 1, 2, \cdots, k, j = 1, 2, \cdots, n$。

一般希望透過變異數分析來了解各組間的期望值($\mu_i = \mu + \tau_i$)是否相同。因此使用變異數分析相當於檢定所有k組資料的期望值是否相等, 亦即

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0, \\ &H_1 : \tau_i \neq 0, \quad \text{至少存在一}i, \quad i = 1, 2, \cdots, k。 \end{aligned} \tag{8.7}$$

針對(8.6)式和(8.7)的假設, 我們可得如表8.1的單因子變異數分析表。

表8.1 單因子變異數分析表

變異原因	平方和	自由度	均方	F_0
組間	$SS_R = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$	$k-1$	$MS_R = \frac{SS_R}{k-1}$	$\frac{MS_R}{MS_E}$
組內	$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$	$k(n-1)$	$MS_E = \frac{SS_E}{k(n-1)}$	
總和	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$kn-1$		

其中$\bar{y}_{i.}$代表第i組中n個觀測值的平均, 即$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$。$\bar{y}_{..}$表示整體平均, 即$\bar{y}_{..} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$。$F_0$為檢定統計量, 且在$H_0$的假設下, F_0服從F分佈, 自由度分別為$k-1$, $k(n-1)$。故若$F_0 > F_{\alpha,k-1,k(n-1)}$ 則拒絕H_0的假設, 反之則接受H_0。

Montgomery(2001)曾研究過關於男性襯衫織布的化學合成物纖維抗拉強度的實驗。此強度是受由纖維中棉花的百分比所影響。資料內容如

表8.2, 圖8.3為其盒鬚圖。根據此筆數據, 可得變異數分析資料如圖8.4。

表8.2 纖維抗拉強度數據資料

棉花百分比	抗拉強度				
15	7	7	15	11	19
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

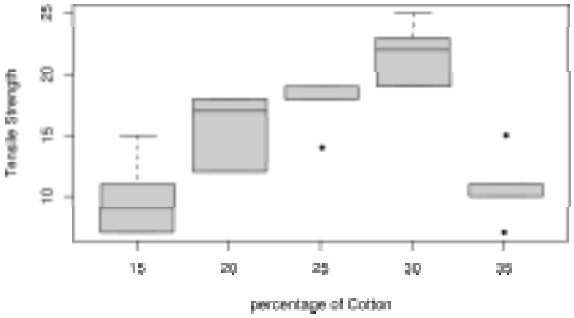


圖 8.3 抗拉強度盒鬚圖

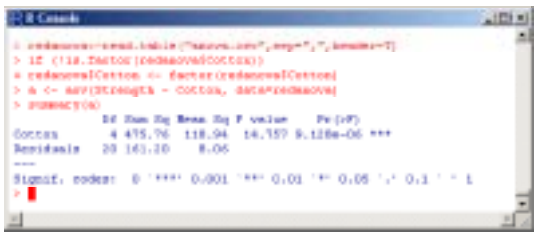


圖 8.4 抗拉強度變異數分析指令及結果

8.3 使用統計軟體R

8.1節 簡單線性迴歸分析

碳氫化合物例子的分析指令。

1. 輸入碳氫化合物的數據資料。

   > hydrocarbon <−read.table("regression.csv",header=F,sep=",")

2. 利用"lsfit"指令作最小平方法的估計。

   > carbon <−lsfit(x =hydrocarbon[,1], y = hydrocarbon[,2])

3. 將估計結果整理並呈現出來。

   >ls.print(carbon)

4. 畫出x和y的散佈圖。

   > plot(hydrocarbon[,1], hydrocarbon[,2],ylab="y",xlab="x ",pch=20)

5. 加入迴歸線。

   > abline(0.2659,0.0016)

   "abline(截距, 斜率)"表示在圖形中增加一條線。

6. 在圖形中標出迴歸函數y=0.2659+0.0016x。

   >text(35,0.3,"y=0.2659+0.0016x")

老忠實噴泉例子的分析指令。

1. 利用"lm"指令配適出老忠實噴泉的線性迴歸模型。

   > fit.faithful <−lm(eruptions .,faithful)

2. 將配適結果整理並呈現出來。

   > summary(fit.faithful)

3. 畫出等待時間(waiting)和噴發時間間隔(eruptions)的散佈圖。

```
> plot(faithful$waiting,faithful$eruptions,pch=20,xlab="waiting",
+ ylab="eruptions",main="The Old Faithful geyser")
```

4. 加入迴歸線。

```
> abline(-1.874016,0.075628)
```

5. 在圖形中標出迴歸函數y=-1874016+0.075628x。

```
> text(82,2.9,"y=-1.874016+0.075628x")
```

6. 給出老忠實噴的變異數分析表

```
> aov.faithful <-aov(lm(eruptions .,faithful))
```

7. 將變異數分析表整理並呈現出來。

```
> summary(aov.faithful)
```

8. 畫出老忠實噴泉的配適值與殘差的散佈圖。

```
> plot(fitted(fit.faithful),residuals(fit.faithful),pch=20,
+ xlab="fitted value",ylab="residuals")
```

9. 加入x=0的水平線。

```
> abline(h = 0, lty = 2, col = "lightgray")
```

10. 畫出老忠實噴泉數據殘差的常態Q-Q圖。

```
> qqnorm(residuals(fit.faithful),pch=20)
> qqline(residuals(fit.faithful))
```

8.2節 變異數分析

1. 輸入抗拉強度的數據資料。

```
> redanova<-read.table("anova.csv",sep=",",header=T)
```

2. 畫出抗拉強度的盒鬚圖。

```
> boxplot(Strength ~ Cotton, data=redanova,col=c(gray(0.8)),
+ xlab="percentage of Cotton",ylab="Tensile Strength",pch=20)
```

3. 如果因子變數本身非R的因子(factor)型態, 先將其轉換成factor。

```
> if (!is.factor(redanova$Cotton))
+ redanova$Cotton <-factor(redanova$Cotton)
```

4. 利用"aov"函數來分析數據資料。

```
> a <-aov(Strength ~ Cotton, data=redanova)
```

5. 將結果整理並呈現出來。

```
> summary(a)
```

習　題

1. Matis and Wehrly(1979)曾研究溫度對翻車魚(sunfish)的存活之影響。在一定溫度下, 經x單位時間, 翻車魚存活y比例的數據如下:

$$(0.10, 1.00), (0.15, 0.95), (0.20, 0.95), (0.25, 0.90), (0.30, 0.85),$$
$$(0.35, 0.70), (0.40, 0.65), (0.45, 0.60), (0.50, 0.55), (0.55, 0.40)。$$

在簡單線性迴歸模型下, 試求β_0及β_1之最小平方估計值。

2. 底下資料為中央氣象局資訊服務網站(http://www.cwb.gov.tw/)從1971至2000年所收集之台北、台中及高雄月平均氣溫。試根據此筆數據給出變異數分析表。

月份	1月	2月	3月	4月	5月	6月
台北	15.8	15.9	18	21.7	24.7	27.4
台中	16.2	16.8	19.4	23	25.7	27.5
高雄	18.8	19.7	22.3	25.2	27.2	28.4

月份	7月	8月	9月	10月	11月	12月
台北	29.2	28.8	27.1	24.3	20.9	17.6
台中	28.5	28	27.2	24.9	21.4	17.8
高雄	28.9	28.3	27.9	26.4	23.4	20.2

第九章

統計品質管制實例分析

9.1 管制圖

管制圖可分為兩大類, 一種為計量管制圖(variable control charts), 另一種為計數管制圖(attributes control charts)。計量管制圖包含了$\overline{X}$管制圖($\overline{X}$ chart)、R管制圖(R chart, 或稱為全距管制圖)及S管制圖(S chart, 或稱為標準差管制圖)。而計數管制圖則有P管制圖(P chart), c管制圖(c chart)及u管制圖(u chart)。

管制圖包含一中心線(central line), 代表製程在管制狀態下產品品質特性的平均數。另外還包含兩條水平線, 分別稱為上管制限(upper control limit, 簡稱UCL)和下管制限(low control limit, 簡稱LCL)。若製程是在管制中, 則管制圖上幾乎所有樣本點皆落在管制上下限之間。底下幾節我們將給出一些例子, 說明何謂管制圖。要利用R來繪製管制圖時, 需先下載名為"qcc"的資料庫。

9.2 計量管制圖

計量(variable)代表品質的特徵可以數字來表示。當要對一個計量

表9.1 活塞環的樣本資料

樣本	內部直徑				
1	74.030	74.002	74.019	73.992	74.008
2	73.995	73.992	74.001	74.011	74.004
3	73.988	74.024	74.021	74.005	74.002
4	74.002	73.996	73.993	74.015	74.009
5	73.992	74.007	74.015	73.989	74.014
6	74.009	73.994	73.997	73.985	73.993
7	73.995	74.006	73.994	74.000	74.005
8	73.985	74.003	73.993	74.015	73.988
9	74.008	73.995	74.009	74.005	74.004
10	73.998	74.000	73.990	74.007	73.995
11	73.994	73.998	73.994	73.995	73.99
12	74.004	74.000	74.007	74.000	73.996
13	73.983	74.002	73.998	73.997	74.012
14	74.006	73.967	73.994	74.000	73.984
15	74.012	74.014	73.998	73.999	74.007
16	74.000	73.984	74.005	73.998	73.996
17	73.994	74.012	73.986	74.005	74.007
18	74.006	74.010	74.018	74.003	74.000
19	73.984	74.002	74.003	74.005	73.997
20	74.000	74.010	74.013	74.020	74.003
21	73.988	74.001	74.009	74.005	73.996
22	74.004	73.999	73.990	74.006	74.009
23	74.010	73.989	73.990	74.009	74.014
24	74.015	74.008	73.993	74.000	74.010
25	73.982	73.984	73.995	74.017	74.013

的品質作管制圖時, 所要監測的是這個品質特性的平均數及變異性。若要管制平均數, 通常使用$\overline{X}$管制圖；要管制過程中的分散或變異性則使用R管制圖, 或者使用S管制圖。本節中, 將以R資料庫中的數據資料"活塞環"(piston rings)來繪製底下的計量管制圖。先簡單描述此筆數據的內容。一部自動發動機的活塞環是從鍛造的過程中製造出來的。這些數據是從一個被認為是在控制中(in control)的過程所取得的。由活塞環的製造過程可量得環的內部直徑, 共取得25個樣本, 每個樣本有五筆資料, 資料內容參考表9.1。接下來我們分成三小節介紹不同計量管制圖的畫法。

9.2.1 $\overline{X}$管制圖

我們先簡單介紹一下$\overline{X}$管制圖背後之統計理論。假設一品質特性遵循常態分佈, 其平均值為μ, 變異數為σ^2。設有n個樣本$x_1, \cdots, x_n$, 其樣本平均為

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}。$$

故$\overline{x}$服從$\mathcal{N}(\mu, \frac{\sigma^2}{n})$, 其樣本平均落在

$$\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.1}$$

和

$$\mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.2}$$

之間的機率為$1-\alpha$。若μ和σ已知, 則(9.1)式和(9.2)式可作為$\overline{X}$管制的上、下界限。通常取$Z_{\alpha/2} = 3$當成$\overline{X}$管制圖的管制界限。

若μ和σ未知時, 以$\bar{\bar{x}}$及$\overline{R}/d_2$分別估計μ及σ, 其理論請參考一般的統計品質管制的書籍, 如Montgomery(1996), 在此不作細部介紹。假設有m個樣本, 每一個樣本有n個觀測值, 則

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m}{m}。$$

以$R_1, \cdots, R_m$分別代表m個樣本的樣本全距, 則

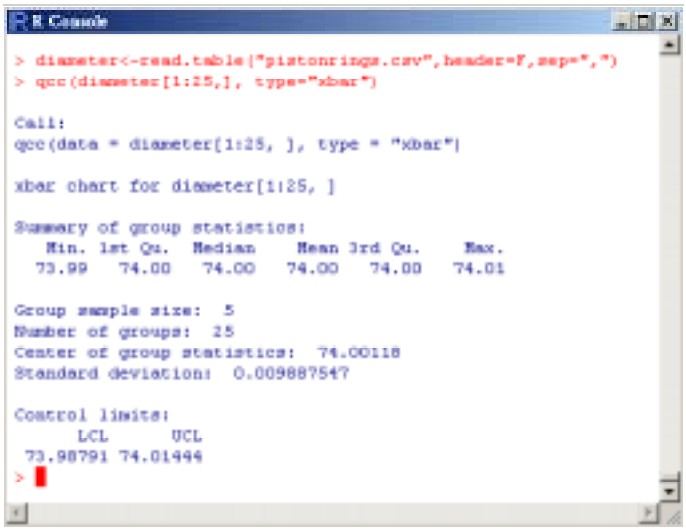
$$\overline{R} = \frac{R_1 + R_2 + \cdots + R_m}{m}。$$

另外σ的估計值與d_2有關, 此d_2值可由統計品質管制的書籍查表得到。

有了上面的估計值, 則$\overline{X}$管制圖的參數分別為

$$\begin{aligned} \text{上管制限}(UCL) &= \bar{\bar{x}} + \frac{3}{d_2\sqrt{n}}\overline{R}, \\ \text{中心線} &= \bar{\bar{x}}, \\ \text{下管制限}(LCL) &= \bar{\bar{x}} - \frac{3}{d_2\sqrt{n}}\overline{R}。 \end{aligned}$$

我們以活塞環的例子來看看$\overline{X}$管制圖的指令及圖形, 參考圖9.1及圖9.2。



```
R Console
> diameter<-read.table("pistonrings.csv",header=F,sep=",")
> qcc(diameter[1:25,], type="xbar")

Call:
qcc(data = diameter[1:25, ], type = "xbar")

xbar chart for diameter[1:25, ]

Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  73.99   74.00   74.00   74.00   74.00   74.01

Group sample size:  5
Number of groups:  25
Center of group statistics:  74.00118
Standard deviation:  0.009887547

Control limits:
      LCL      UCL
 73.98791 74.01444
>
```

圖 9.1 活塞環的$\overline{X}$管制圖指令

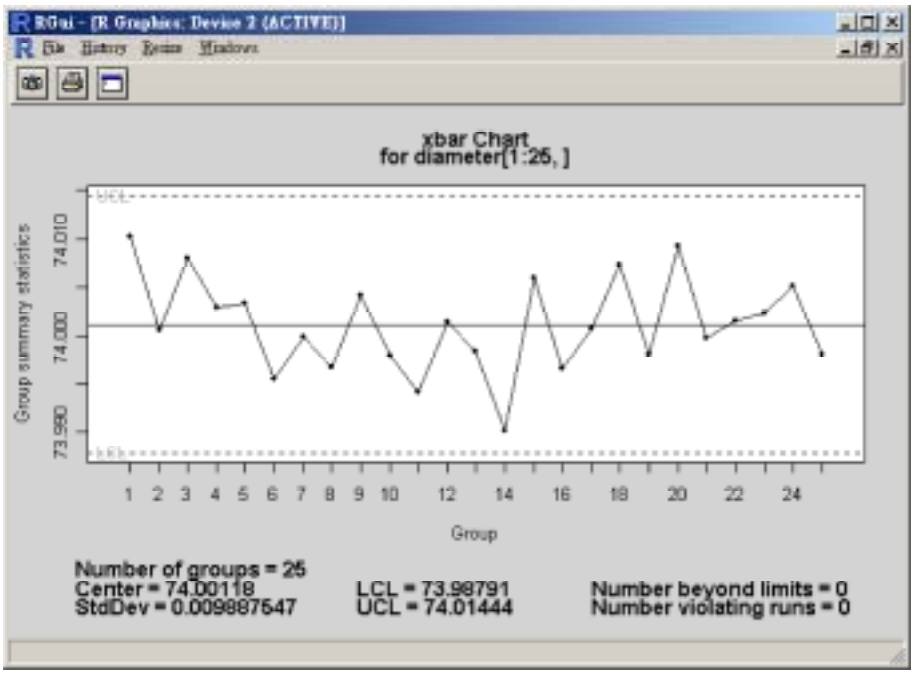


圖 9.2 活塞環的$\overline{X}$管制圖

9.2.2 R管制圖

由前一小節知樣本全距可用來估計製程的標準差, 因此, 製程的變異可以R值繪製出來, 稱此管制圖為R管制圖。R管制圖的中心線以$\overline{R}$表之, 底下僅給出R管制圖的參數為

$$
\begin{aligned}
\text{上管制限}(UCL) &= \overline{R} + 3\frac{d_3\overline{R}}{d_2}, \\
\text{中心線} &= \overline{R}, \\
\text{下管制限}(LCL) &= \overline{R} - 3\frac{d_3\overline{R}}{d_2}。
\end{aligned}
$$

其中d_2及d_3可由統計品質管制的書籍查表得到。接下來看活塞環之R管制圖的指令及圖形, 參考圖9.3及圖9.4。

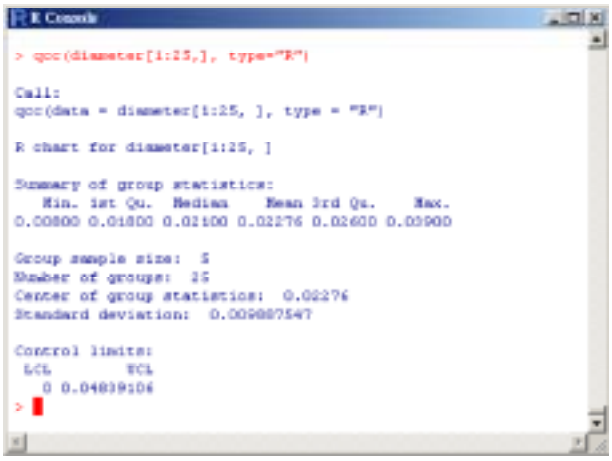


圖 9.3 活塞環的R管制圖指令

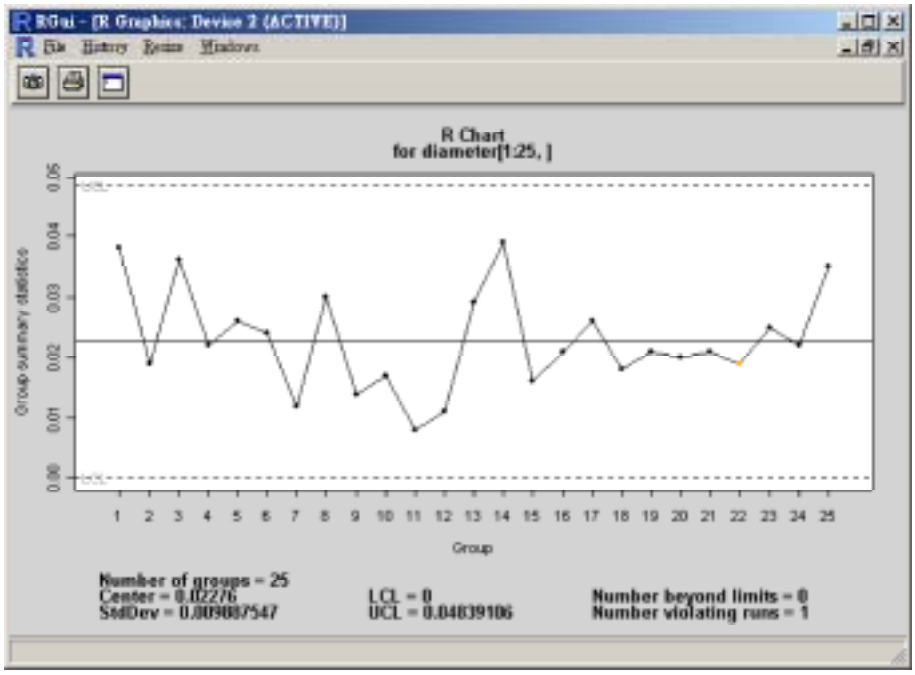


圖 9.4 活塞環的R管制圖

9.2.3 S管制圖

假如σ^2是未知的, 則σ^2的不偏估計值為

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

但s並非σ的不偏估計值, 如果在常態分佈下, 可用s/c_4估計σ, 其中$c_4 = (\frac{2}{n-1})^{1/2}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$。以$\overline{S}$代表樣本標準差的平均數, 即為

$$\overline{S} = \frac{1}{m}\sum_{i=1}^{m} s_i。$$

則S管制圖的參數為

$$\begin{aligned} \text{上管制限}(UCL) &= \overline{S} + 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}, \\ \text{中心線} &= \overline{S}, \\ \text{下管制限}(LCL) &= \overline{S} - 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}。 \end{aligned}$$

其中c_4亦可由表查得。圖9.5及圖9.6為活塞環之S管制圖的指令及圖形。

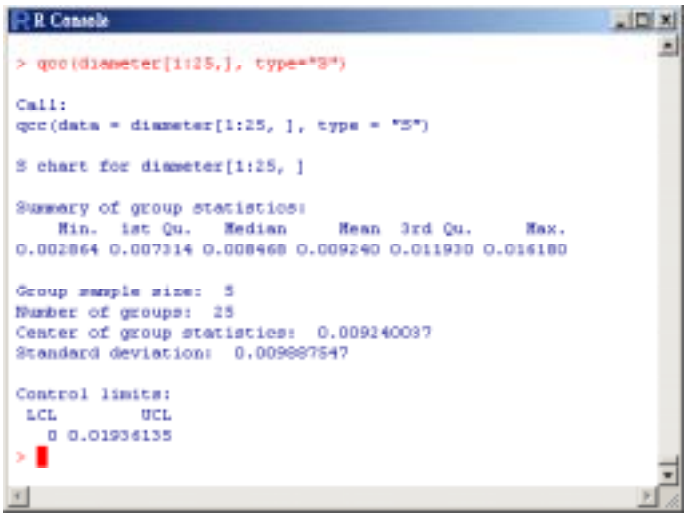


圖 9.5 活塞環的S管制圖指令

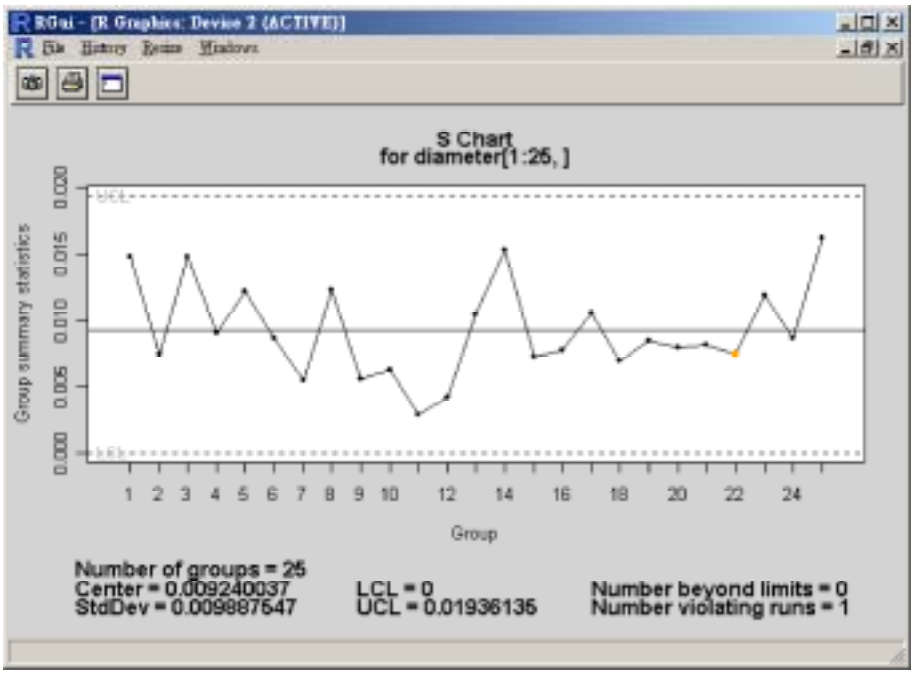


圖 9.6 活塞環的S管制圖

9.3 計數管制圖

有些時候品質特性並不能簡單的用數字來表現, 通常依是否符合規格的品質特性, 來判斷被檢驗的項目是否為良品。以合格或不合格來作為品質特性的分類依據, 稱之為計數值(attribute)。我們介紹三種計數值管制圖, 分別是p管制圖, c管制圖和u管制圖。探討關於製造過程中產生的不合格或不良的產品比率, 稱為不合格率管制圖或不良率管制圖(control chart for fraction nonconforming), 或稱為p管制圖。另外在某些情況下, 採用缺點數或不合格數較不合格率簡單, 此種情況採用缺點數管制圖(control chart for nonconformities), 或稱為c管制圖。最後一種稱為單位缺點管制圖(control chart for nonconformities per unit), 或稱為u管制圖, 是以在每單位平均缺點數作為製程管制圖的一個依據。

9.3.1 p管制圖

設製程中眞正不合格率為p, p未知時, 由數據資料可估計p。設有m個樣本, 每個樣本有n個觀察值。若樣本i有D_i個不合格品, 則第i個樣本的不合格率為

$$\hat{p}_i = \frac{D_i}{n} \quad i = 1, 2, \cdots, m,$$

則這些樣本不合格率的平均數為

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{mn} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}。$$

未知的不合格率以$\bar{p}$來估計, 則不合格率管制圖的中心線和管制界限為

$$\begin{aligned} \text{上管制限}(UCL) &= \bar{p} + 3\frac{\sqrt{\bar{p}(1-\bar{p})}}{n}, \\ \text{中心線} &= \bar{p}, \\ \text{下管制限}(LCL) &= \bar{p} - 3\frac{\sqrt{\bar{p}(1-\bar{p})}}{n}。 \end{aligned}$$

以R資料庫中的數據資料“柳橙汁”(orange juice)的前30筆資料為例, 資料內容參考表9.2。此筆資料是紀錄裝盛冷凍濃縮柳橙汁時, 是否會漏出盛裝物, 來看是否符合品管規定。因冷凍柳橙汁是用6盎司的紙罐裝, 這種紙罐是先用紙板製成罐狀, 然後底部加裝金屬板而成。所以紙罐的品質檢驗, 是當紙罐裝滿液體後, 檢查液體是否會從側邊或底部的接縫漏出, 即紙罐是否符合品質管制, 是指側邊或底部的接縫是否適當, 故利用管制圖來監測此部機器的生產。因此根據此筆資料可繪製樣本的p管制圖, 如圖9.7及圖9.8。

表9.2 柳橙汁的樣本資料

樣本編號	不合格數	樣本不合格率	樣本編號	不合格數	樣本不合格率
1	12	0.24	16	8	0.16
2	15	0.3	17	10	0.2
3	8	0.16	18	5	0.1
4	10	0.2	19	13	0.26
5	4	0.08	20	11	0.22
6	7	0.14	21	20	0.4
7	16	0.32	22	18	0.36
8	9	0.18	23	24	0.48
9	14	0.28	24	15	0.3
10	10	0.2	25	9	0.18
11	5	0.1	26	12	0.24
12	6	0.12	27	7	0.14
13	17	0.34	28	13	0.26
14	12	0.24	29	9	0.18
15	22	0.44	30	6	0.12

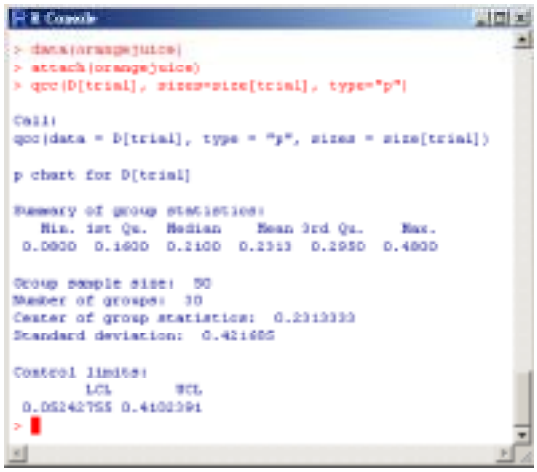


圖 9.7 柳橙汁的p管制圖指令

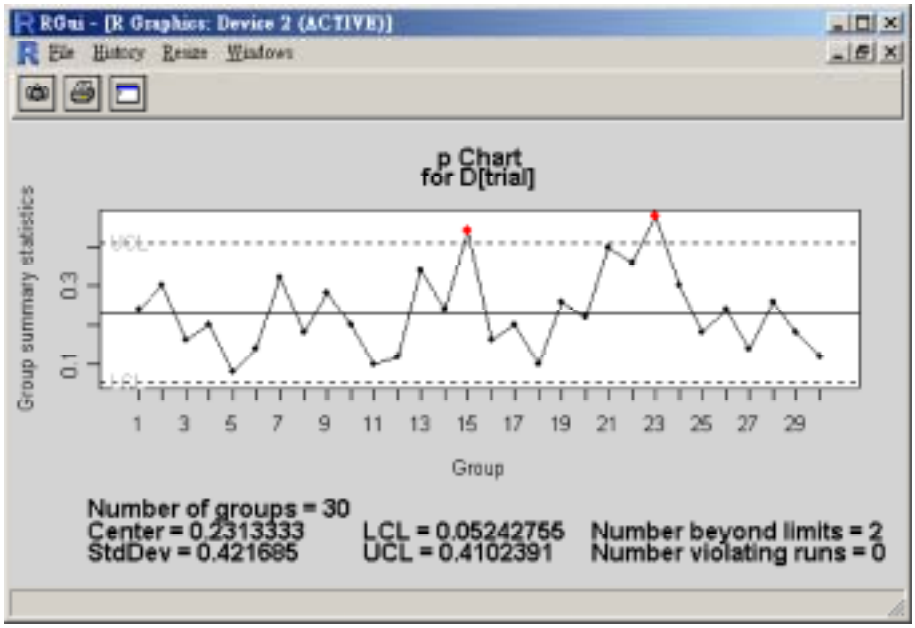


圖 9.8 柳橙汁的p管制圖

9.3.2 c管制圖

在此僅介紹樣本數相同的c管制圖, 其他相關的內容請參考統計品質管制的書籍。c管制圖的理論假設是假設檢驗單位發生缺點是依據波松分佈, 參數為c, 即為

$$p(x) = \frac{e^{-c}c^x}{x!} \quad x = 0, 1, 2, \cdots,$$

其中x是缺點數。若c未知, 以樣本檢驗單位中觀測值的平均數來估計, 則可得c管制圖的參數為

$$\begin{aligned} \text{上管制限}(UCL) &= \bar{c} + 3\sqrt{\bar{c}}, \\ \text{中心線} &= \bar{c}, \\ \text{下管制限}(LCL) &= \bar{c} - 3\sqrt{\bar{c}}。 \end{aligned}$$

以R資料庫中的“印刷電路板”為例子。數據資料如表9.3, 紀錄26個連續樣本中所觀測到的不良品數, 每一個樣本取100個印刷電路板。由數據資料

可得其c管制圖指令及圖形, 如圖9.9及圖9.10。

表9.3 印刷電路板的樣本資料

樣本編號	不良品數	樣本編號	不良品數
1	21	14	19
2	24	15	10
3	16	16	17
4	12	17	13
5	15	18	22
6	5	19	18
7	28	20	39
8	20	21	30
9	31	22	24
10	25	23	16
11	20	24	19
12	24	25	17
13	16	26	15

9.3.3 u管制圖

u管制圖在此是以每一檢驗單位的缺點平均數來建立管制圖。若一個樣本有n個檢驗單位, 共有c個缺點, 則每一檢驗單位的缺點平均數是$u = \frac{c}{n}$。則u管制圖的參數為

$$\begin{aligned} \text{上管制限}(UCL) &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}}, \\ \text{中心線} &= \bar{u}, \\ \text{下管制限}(LCL) &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}}。 \end{aligned}$$

```
R Console
> data(circuit)
> attach(circuit)
> qcc(x[trial], sizes=size[trial], type="c")

Call:
qcc(data = x[trial], type = "c", sizes = size[trial])

c chart for x[trial]

Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   16.00   19.00   19.85   24.00   39.00

Group sample size:  100
Number of groups:  26
Center of group statistics:  19.84615
Standard deviation:  4.454902

Control limits:
      LCL      UCL
 6.481447 33.21086
>
```

圖 9.9 印刷電路板的c管制圖指令

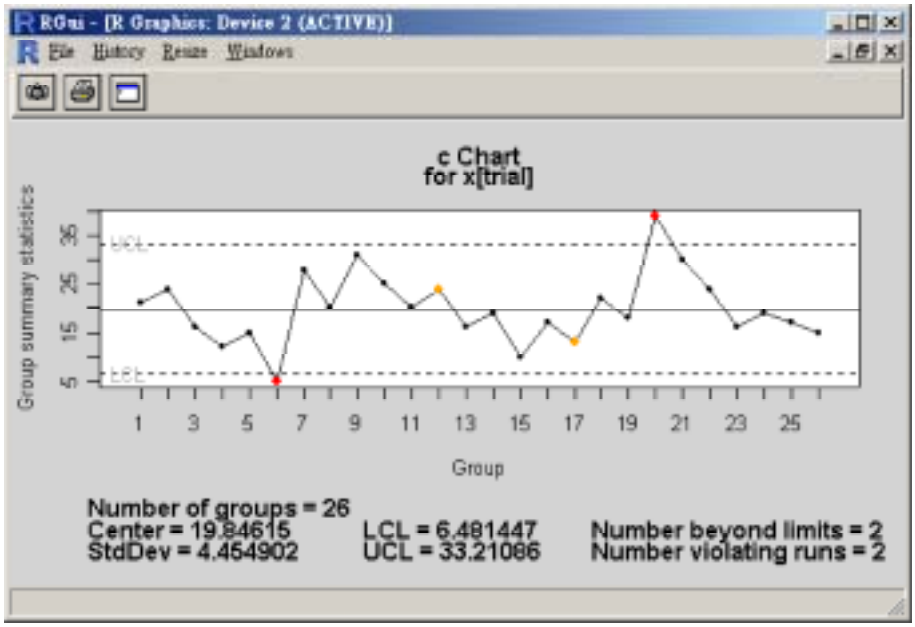


圖 9.10 印刷電路板的c管制圖

以R資料庫中的“個人電腦製造商”(pcmanufact)為例子。一個人電腦的製造商希望在每一條生產線上建立u管制圖，樣本大小為5部電腦。20個樣本的缺點數如表9.4，由數據資料可得u管制圖指令及圖形，如圖9.11及圖9.12。

表9.4 個人電腦製造商的樣本資料

樣本編號	缺點總數	樣本大小	樣本編號	缺點總數	樣本大小
1	10	5	11	9	5
2	12	5	12	5	5
3	8	5	13	7	5
4	14	5	14	11	5
5	10	5	15	12	5
6	16	5	16	6	5
7	11	5	17	8	5
8	7	5	18	10	5
9	10	5	19	7	5
10	15	5	20	5	5

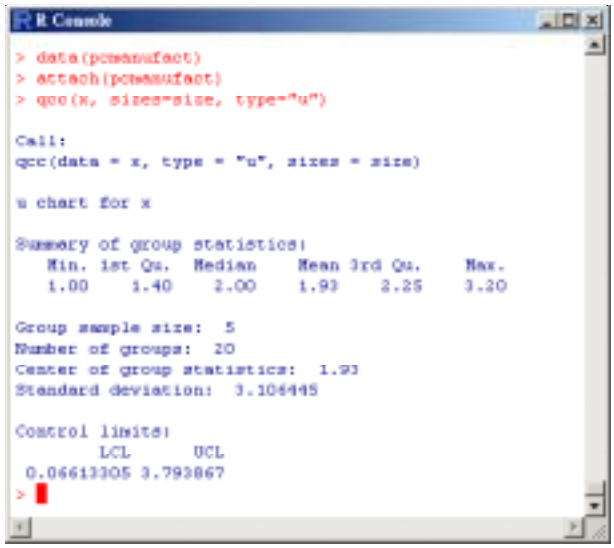


圖 9.11 個人電腦製造商的u管制圖指令

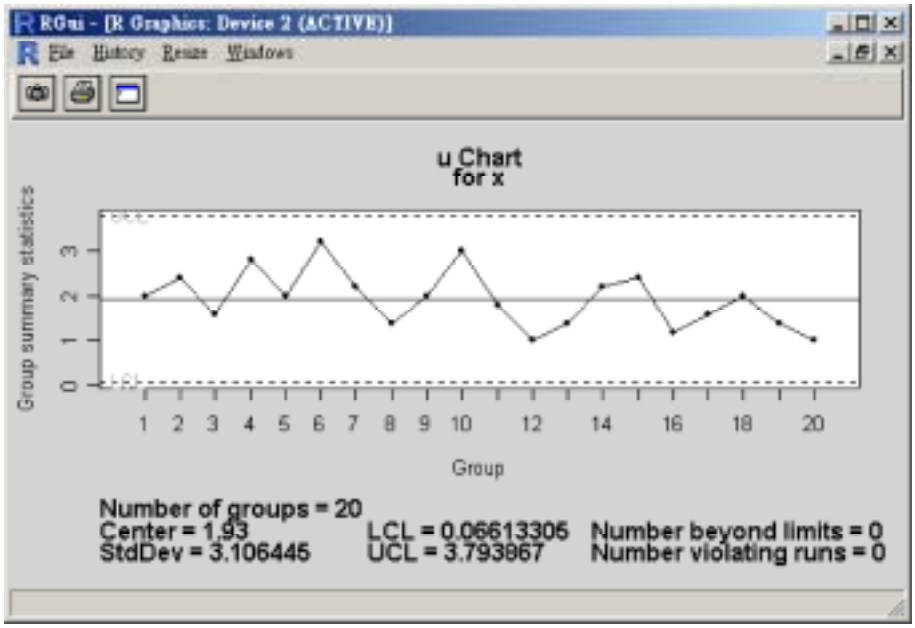


圖 9.12 個人電腦製造商的u管制圖

9.4 累積和及指數加權移動平均管制圖

在此節我們另外給出兩個統計品質管制上會用到的管制圖。一稱為累積和管制圖(cumulative-sum control chart, 或簡稱cusum), 另一種為指數加權移動平均管制圖(esponentially weighted moving-average control chart, 或簡稱EWMA)

以前面提過之活塞環的例子, 利用R的指令“cusum”及“ewma”可分別畫出累積和管制圖及指數加權移動平均管制圖, 如圖9.13及圖9.14。相關的理論背景請參考一般的統計品質管制書籍, 如Montgomery(1996)。

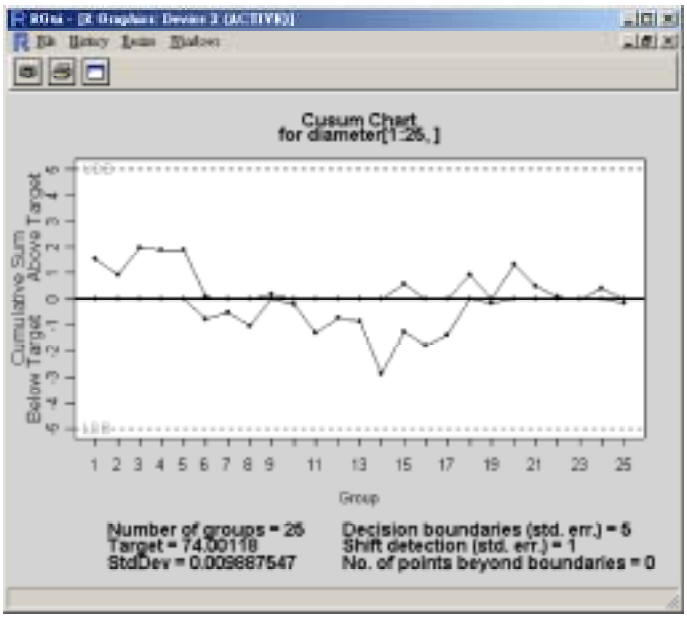


圖 9.13 活塞環的累積和管制圖

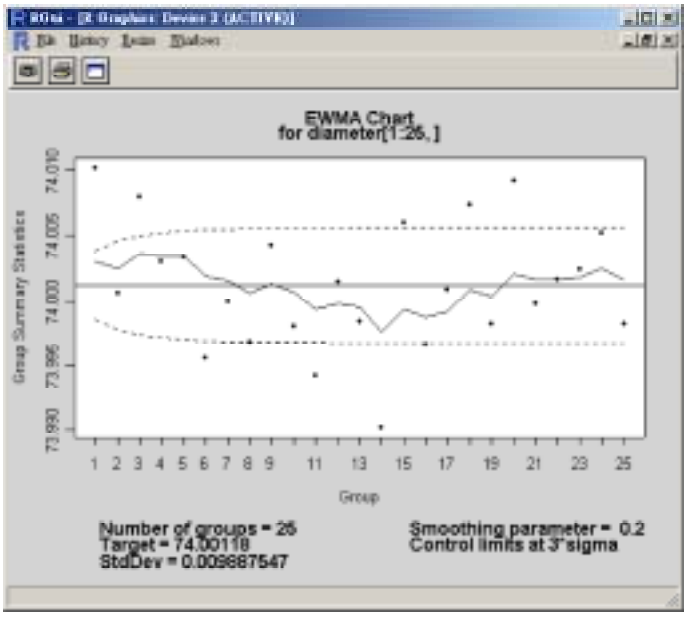


圖 9.14 活塞環的指數加權移動平均管制圖

9.5 使用統計軟體R

9.2節 計量管制圖

1. 輸入活塞環的數據資料。

   > diameter<-read.table("pstonrings.csv",header=T,sep=",")

2. 活塞環的$\overline{X}$管制圖。

   > qcc(diameter[1:25], type="xbar")

3. 畫出活塞環的R管制圖。

   > qcc(diameter[1:25], type="R")

4. 畫出活塞環的S管制圖。

   > qcc(diameter[1:25], type="S")

9.3節 計數管制圖

1. 畫出柳橙汁的p管制圖。

   > data(orangejuice)

   > attach(orangejuice)

   > qcc(D[trial], sizes=size[trial], type="p")

2. 畫出印刷電路板的c管制圖。

   > data(circuit)

   > attach(circuit)

   > qcc(x[trial], sizes=size[trial], type="c")

3. 畫出個人電腦製造商的u管制圖。

   > data(pcmanufact)

   > attach(pcmanufact)

   > qcc(x, sizes=size, type="u")

9.4節 累積和及指數加權移動平均管制圖

1. 畫出活塞環的累積和管制圖。

```
> q <-qcc(diameter[1:25,], type="xbar", nsigmas=3, plot=FALSE)
> cusum(q)
```

2. 畫出活塞環的指數加權移動平均管制圖。

```
> ewma(q, lambda=0.2)
```

習　題

1. 試利用資料庫中名為"dyedcloth"的數據來建立c管制圖及u管制圖。此資料是一家紡織工廠的數據。以每50平方公尺的缺點數來檢查染色過的布,試利用此資料來建立c管制圖及u管制圖。

2. Montgomery(1991)給出如下表之數據。高壓電廠電壓的輸出值須為350V,若每天取四個單位為樣本並檢查以確保製程在管制之下。下列資料x_i為第i個觀測值-350之後再乘以10。試利用此筆資料建立$\overline{X}$及R管制圖。

編號	x_1	x_2	x_3	x_4
1	6	9	10	15
2	10	4	6	11
3	7	8	10	5
4	8	9	6	13
5	9	10	7	13
6	12	11	10	10
7	16	10	8	9
8	7	5	10	4
9	9	7	8	12
10	15	16	10	13

編號	x_1	x_2	x_3	x_4
11	8	12	14	16
12	6	13	9	11
13	16	9	13	15
14	7	13	10	12
15	11	7	10	16
16	15	10	11	14
17	9	8	12	10
18	15	7	10	11
19	8	6	9	12
20	14	15	12	16

第十章

時間序列實例分析

10.1 時間序列模型建立步驟

在這一章中, 我們並不詳細介紹時間序列(time series)每一個模型背後的理論架構, 而是利用時間序列上著名的一些例子, 說明時間序列數據的分析過程。底下我們簡單說明模型建立的一些步驟。

在此, 給出一般的ARIMA(p,d,q)模型

$$(1-\phi_1 B-\cdots-\phi_p B^p)(1-B)^d Z_t=\theta_0+(1-\theta_1 B-\cdots-\theta_q B^q)a_t。$$

在模型中, $\phi_i, i=1,\cdots,p$為AR的參數, 位階(order)為p。$\theta_j, j=1,\cdots,q$為MA的參數, 位階為q。d決定模型須經過幾次的差分動作, Z_t則為在時間t得到的觀察值, a_t稱為白噪音(white noise), 一般取a_t滿足$E(a_t)=0$, $Var(a_t)$為一常數。在給定一時間序列資料, 我們通常使用下列步驟來決定一可能的模型。

1. 先繪製時間序列圖, 並且選擇適當的轉換方式。對任何的時間序列分析, 首先採取的步驟皆是先畫出資料的圖形。透過圖形的檢驗, 可提供關於資料的一些資訊, 像是趨勢、週期、離群值或者變異等。對時間序列資料有時需要經過變異數平穩的轉換或者差分。常用的轉換如對資料取log或開平方。

2. 計算並檢驗原始資料的ACF(autocorrelation function)及PACF(partial autocorrelation function)圖形。由資料的ACF及PACF圖形通常可推估p, q及d值。

3. 當$d > 0$, 檢定所決定的趨勢項θ_0。

要利用R來分析時間序列資料時, 需先下載名為"ts"的資料庫。

10.2 加拿大山貓數量資料分析

在R的資料庫中加拿大山貓(lynx)數量的數據資料名為"lynx", 此筆資料記錄加拿大西北部馬更些(Mackenzie)流域附近每年的山貓數量, 資料所收集的時間是西元1821至1934年。利用"plot"指令先畫出加拿大山貓數量的時間序列圖, 如圖10.1。圖10.2為加拿大山貓數量的ACF及PACF圖。

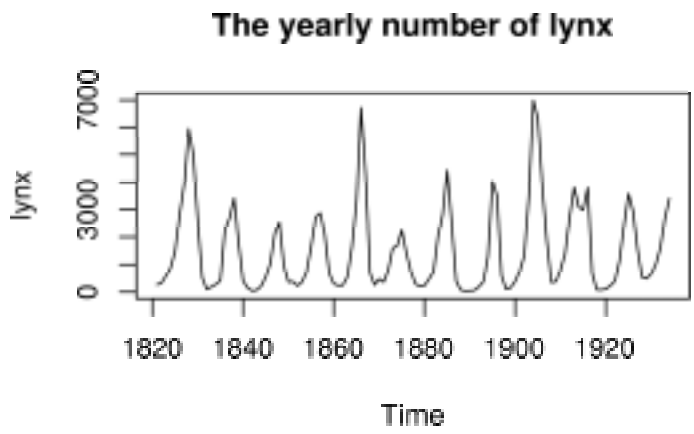


圖 10.1 加拿大山貓數量的時間序列圖

由圖10.3的常態Q-Q圖來看, 資料並不是很符合常態分佈, 因此對原始資料作對數轉換。轉換後的資料由圖10.4的常態Q-Q圖可明顯看出較原始資料更符合常態分佈。對經過對數轉換後的數據, 利用"plot"、"acf"及"pacf"指

令分別畫出時間序列圖、ACF及PACF圖, 如圖10.5及圖10.6。

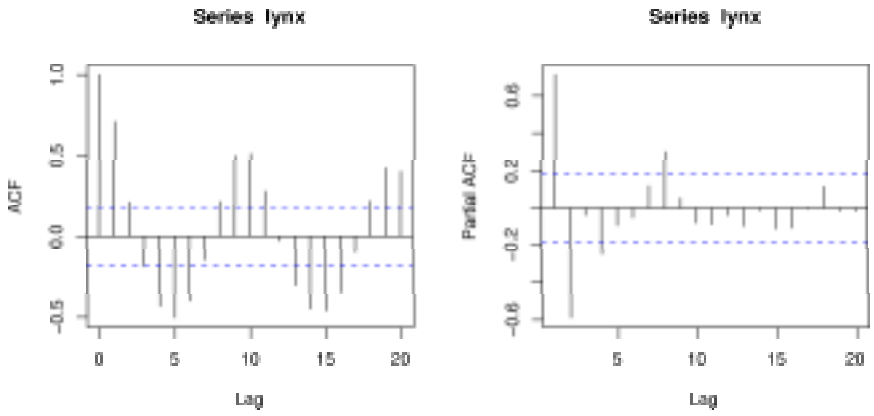


圖 10.2 加拿大山貓數量的ACF及PACF圖

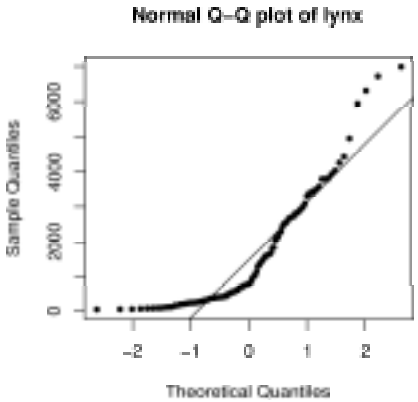


圖 10.3 加拿大山貓數量的常態Q-Q圖

發現經過轉換後之數據仍不平穩, 由圖10.5的序列圖及圖10.6的ACF圖形可粗略看出其具有週期性, 且週期大約在9至10之間。因為ACF呈現指數下降, PACF在lag=2或11之後大致都落在水平線之間。經由這些圖形的判斷, 我們決定配適一個AR(p)的模型。利用R中的指令“ar”可得圖10.7的結果, R自動配適出AR(11)的模型, 並給出每一係數值。

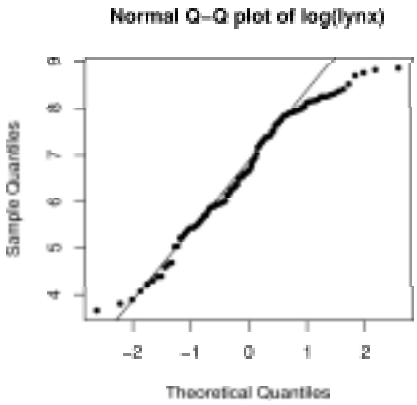


圖 10.4 經對數轉換後, 加拿大山貓數量的常態Q-Q圖

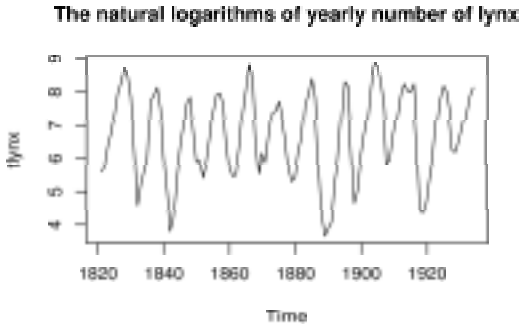


圖 10.5 經對數轉換後, 加拿大山貓數量的時間序列圖

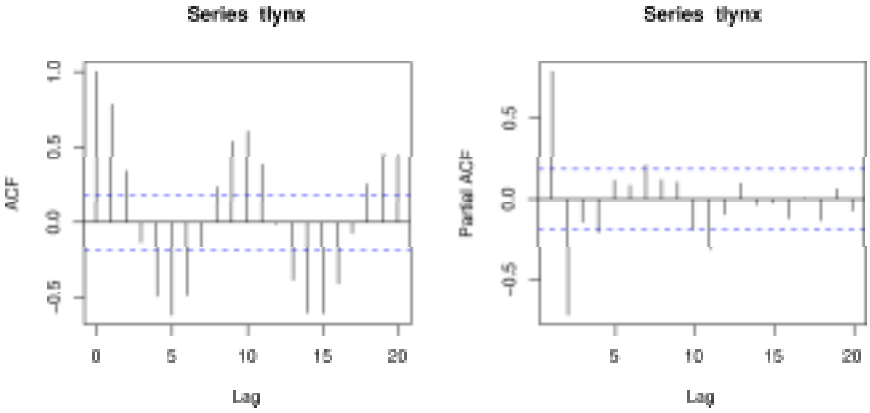
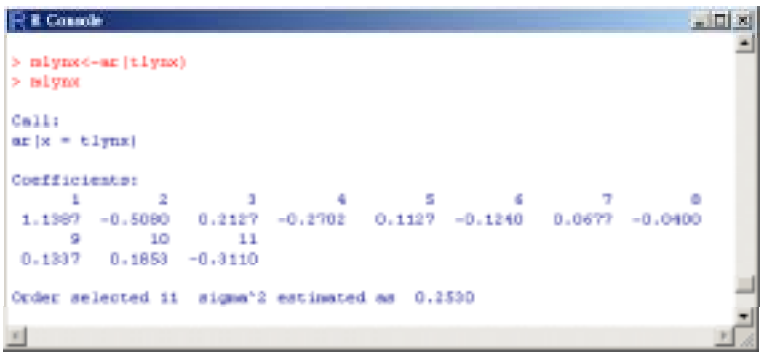


圖 10.6 經對數轉換後, 加拿大山貓數量的ACF及PACF圖



```
> mlynx<-ar(tlynx)
> mlynx

Call:
ar(x = tlynx)

Coefficients:
      1       2       3       4       5       6       7       8
 1.1987 -0.5080  0.2127 -0.2702  0.1127 -0.1240  0.0677 -0.0400
      9      10      11
 0.1337  0.1853 -0.3110

Order selected 11  sigma^2 estimated as  0.2530
```

圖 10.7 經對數轉換後, 加拿大山貓數量的模型配適結果

10.3 太陽黑子數據資料分析

在R的資料庫中太陽黑子的數據資料名為“sunspots”, 此筆資料記錄從西元1749到1983年間太陽黑子之月及年平均數, 我們僅針對太陽黑子的年平均數作討論, 並設此筆資料為Z_t, t表時間。先畫出其時間序列圖, 如圖10.8。利用R的指令“acf”及“pacf”可分別畫出太陽黑子的ACF及PACF圖形, 如圖10.9。

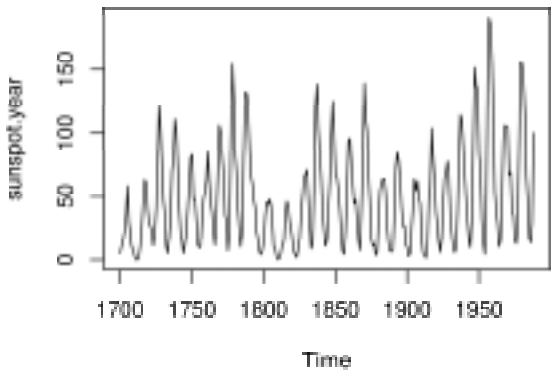


圖 10.8 太陽黑子的時間序列圖

由圖10.8的時間序列圖觀察到變異數可能不為一常數值, 所以考慮將此筆資料做開方轉換。分別畫原始資料Z_t和$\sqrt{Z_t}$的常態Q-Q圖, 如圖10.10。經轉換後之太陽黑子資料的時間序列圖如圖10.11。對轉換後之數據畫出ACF及PACF圖, 參考圖10.12。

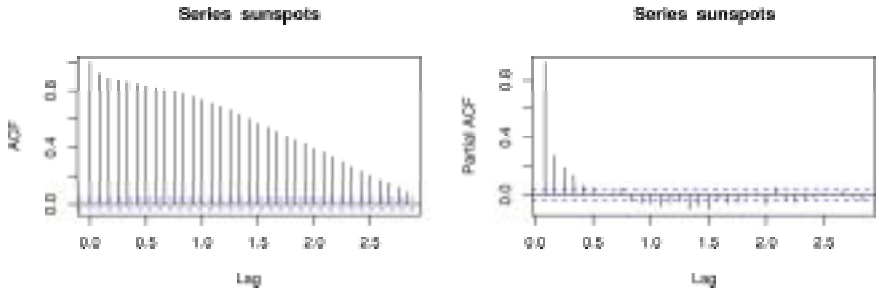


圖 10.9 太陽黑子的ACF及PACF圖

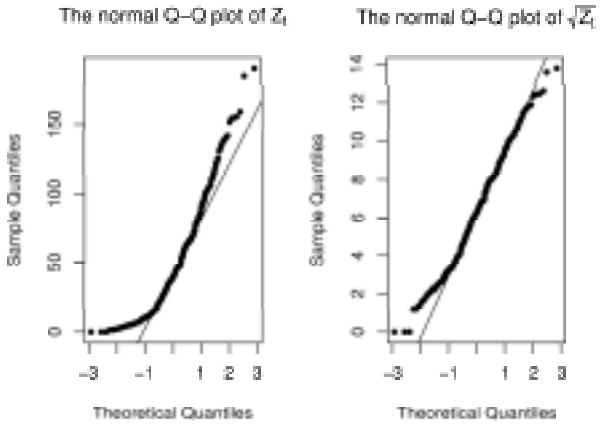


圖 10.10 太陽黑子原始資料Z_t和$\sqrt{Z_t}$的常態Q-Q圖

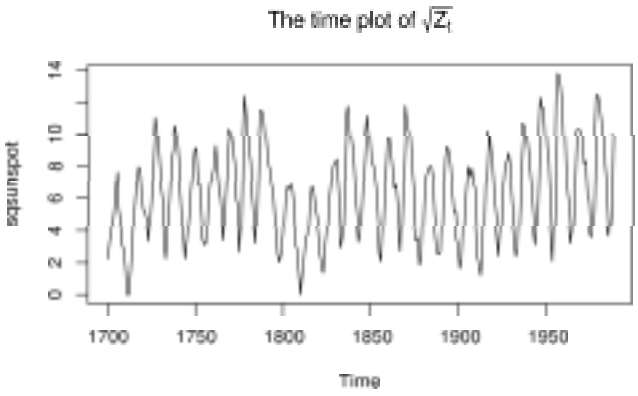


圖 10.11 經開方轉換後, 太陽黑子時間序列圖

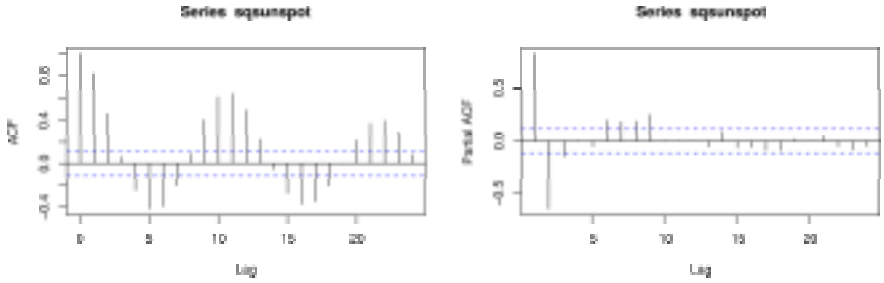


圖 10.12 經開方轉換後, 太陽黑子的ACF及PACF圖

許多學者已試過AR(2)模型和AR(9)模型, 我們同樣利用此二模型來配適轉換後之數據。利用“ar”指令可分別得到AR(2)和AR(9)參數估計值, 如圖10.13及圖10.14。

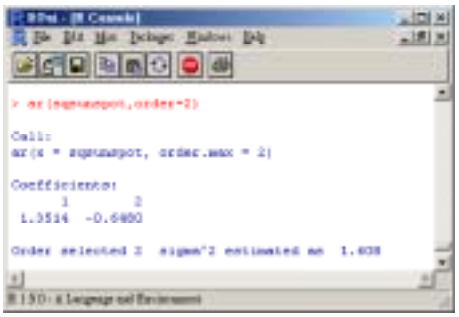


圖 10.13 經開方轉換後, 太陽黑子AR(2)模型配適結果

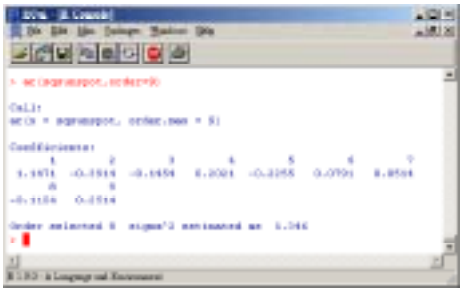


圖 10.14 經開方轉換後, 太陽黑子AR(9)模型配適結果

由圖10.12之ACF可約略看出資料具有週期性, 在一些文獻探討中, 提出週期可能為11。故先將資料去除週期性, 利用“diff”指令對資料作差分, 差分後可畫出時間序列圖、ACF及PACF圖, 如圖10.15及圖10.16。最後考慮ARIMA(2,0,0)×(1,1,0)$_{11}$模型, 利用“arima”指令來配適太陽黑子資料, 結果如圖10.17。

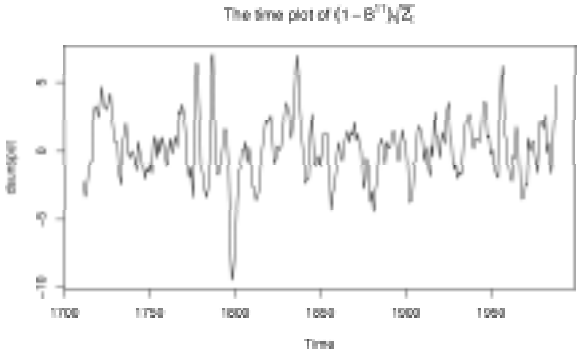


圖 10.15 經差分後, 太陽黑子時間序列圖

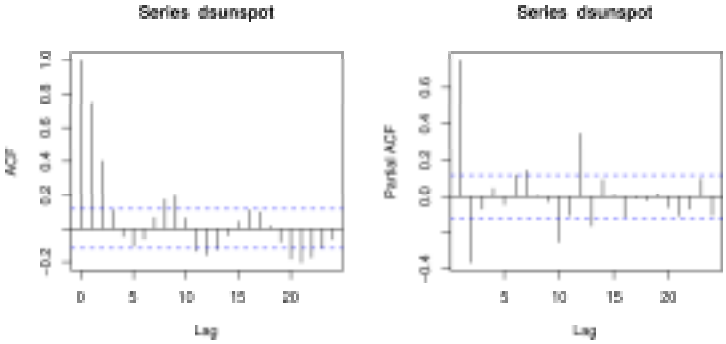


圖 10.16 經差分後, 太陽黑子的ACF及PACF圖

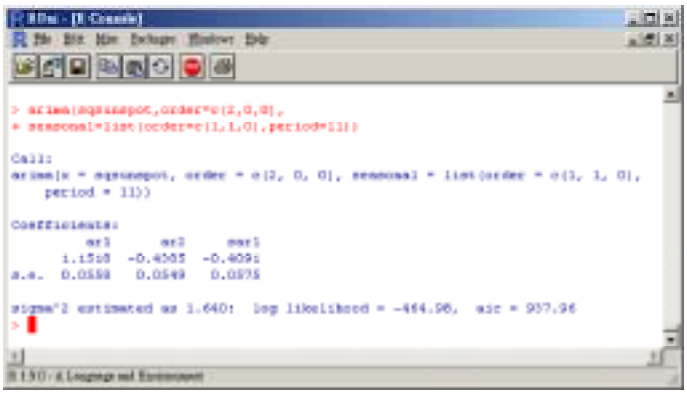


圖 10.17 太陽黑子的模型配適結果

10.4 使用統計軟體R

10.2節 加拿大山貓數量資料分析

1. 輸入資料, 並畫出時間序列圖、ACF、PACF及常態Q-Q圖。

```
> data(lynx)
> plot(lynx,main="The yearly number of lynx")
> acf(lynx)
> pacf(lynx)
> qqnorm(lynx,pch=20,main="Normal Q-Q plot of lynx")
> qqline(lynx)
```

2. 對資料作對數轉換, 並畫出常態Q-Q圖、時間序列圖、ACF及PACF圖。

```
> tlynx <-log(lynx)
> qqnorm(tlynx,pch=20,main="Normal Q-Q plot of log(lynx)")
> qqline(tlynx)
> plot(tlynx,main="The natural logarithms
+ of yearly number of lynx")
> acf(tlynx)
> pacf(tlynx)
```

3. 配適AR(p)模型。

```
> mlynx <-ar(tlynx)
```

10.3節 太陽黑子數據資料分析

1. 先輸入資料, 並畫出時間序列圖、ACF及PACF圖。

```
> data(sunspot)
> plot(sunspot.year,pch=20)
```

```
> acf(sunspot.year)
> pacf(sunspot.year)
```

2. 將資料作開方轉換。分別畫出原始資料及開方後資料的常態Q-Q圖。

```
> sqsunspot<-sqrt(sunspot.year)
> par(mfrow=c(1,2))
> qqnorm(sunspot.year,pch=20,
+ main=expression(paste("The normal Q-Q plot ",Z[t])))
> qqline(sunspot.year)
> qqnorm(sqsunspot,pch=20,
+ main=expression(paste("The normal Q-Q plot ",sqrt(Z[t]))))
> qqline(sqsunspot)
```

3. 畫出開方轉換後之資料的時間序列圖、ACF及PACF圖。

```
> par(mfrow=c(1,1))
> plot(sqsunspot,main=expression(paste("The time plot of ",
+ sqrt(Z[t]))))
> acf(sqsunspot)
> pacf(sqsunspot)
```

4. 配適AR(2)模型。

```
> ar(sqsunspot,order=c(2))
```

5. 配適AR(9)模型。

```
> ar(sqsunspot,order=c(9))
```

6. 對轉換後之資料作差分。並畫出時間序列圖、ACF及PACF。

```
> dsunspot<-diff(sqsunspot,lag=11)
```

```
> plot(dsunspot,pch=20,main=expression(paste("The time plot
+ of ",(1-B^11)sqrt(Z[t]))))
> acf(dsunspot)
> pacf(dsunspot)
```

7. 配適ARIMA(2,0,0)× $(1,1,0)_{11}$模型。

```
> arima(sqsunspot,order=c(2,0,0),
+ seasonal=list(order=c(1,1,0),period=11))
```

習　題

1. 利用資料庫中名為"LakeHuron"的數據,給出此筆數據之時間序列圖、ACF、PACF、常態Q-Q圖。
2. 數據同上,由上題之結果給出一適當的模型

參考文獻

1. 黃文璋 (2003). 數理統計. 華泰文化事業股份有限公司, 台北市.

2. Casella, G. and Berger, R.L. (2002). *Statistical Inference,* 2nd ed. Wadsworth and Brooks/Cole Publishing Company, Pacific Grove, California.

3. Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd ed. John Wiley & Sons Inc., New York.

4. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*, 3rd ed. John Wiley & Sons Inc., New York.

5. Montgomery,D.C. (1996). *Introduction to Statistical Quality Control*, 3rd ed. John Wiley & Sons Inc., New York.

6. Matis, J.H. and Wehrly, T.E. (1979). Stochastic models of compartmental systems. *Biometrics*, 35, 199-220.

7. McDonald, G.C. and Studden, W.J. (1990). Design aspects of regression-based ratio estimation. *Technometrics*, 32, 417-424.

8. Dalgaard, P. (2002). *Introductory Statistics With R.* Springer Verlag, New York.

9. Wei, W.W.S. (1990). *Time Series Analysis : Univariate and Multivariate Methods.* Addison-Wesley Pubilishing Company Inc., New York.

參考網站

1. R的首頁：http://www.r-project.org。
2. 淡江大學統計學系的Net-Stat：http://tkustat.stat.tku.edu.tw。
3. 政治大學統計系余清祥老師個人網站：http://csyue.nccu.edu.tw。
4. 統計製程遠距教學系統：http://spcedu.iem.yzu.edu.tw。

附　表 A

附表A.1 基本操作指令
附表A.2 運算子
附表A.3 資料型態
附表A.4 資料的輸入與輸出
附表A.4a read.table的其他指令
附表A.4b write.table的其他指令
附表A.5 圖形
附表A.5a 其他圖形指令
附表A.6 數學上的函數
附表A.7 統計上的函數
附表A.8 常態分佈
附表A.8a 其他分佈
附表A.9 統計方法
附表A.10 函數及遞迴指令
附表A.10a 函數及遞迴指令的例子

附　表 B

附表B.1 台北、高雄及台中之月平均溫度數據資料
附表B.2 92年翠水庫日平均水位數據資料

附　表 A

附表A.1 基本操作指令

指令	說明
+	未完成指令輸入
<-	設定變數
->	指定給右邊變數
ls(), ls.str(), objects()	查詢
rm()	刪除
help(), help.search(), help.start()	說明
q()	離開R

附表A.2 運算子

指令	說明
+, −, *, /	加,減,乘,除
==	相等
!=	不等
> , <	大於,小於
>= , <=	大於等於,小於等於
&	且(邏輯判斷)
|	或(邏輯判斷)
!	非(邏輯判斷)

附表A.3 資料型態

指令	說明
seq(-5,5,1)	產生-5,-4,-3,$\cdots$,3,4,5
c(1,2,3,4,5)	向量(1,2,3,4,5)
rep(1,5)	重複五個1
as.matrix(x)	x轉換成矩陣型態
as.vector(x)	x轉換成向量型態
as.numeric(x)	x轉換成數字型態
as.character(x)	x轉換成字元型態
factor(x)	x轉換成因子型態
data.frame(data1=c(1,2,3), data2=c(100,200,300))	資料表
attach(x)	將x放入搜尋路徑中
detach(x)	將x從路徑中移除
is.matrix(x)	判斷x是否為矩陣型態
is.vector(x)	判斷x是否為向量型態
is.numeric(x)	判斷x是否為數字型態
is.chacter(x)	判斷x是否為字元型態

附表A.4 資料的輸入與輸出

指令	說明
data(名稱)	讀取內建資料庫
read.table("檔案名稱")	輸入檔案
read.csv("檔案名稱")	輸入以逗號分隔之檔案
read.csv2("檔案名稱")	輸入以分號分隔之檔案
write.table(欲輸出之變數名稱,"輸出之檔案名稱")	輸出檔案

附表A.4a read.table的其他指令

指令	說明
header=TRUE	讀取時, 是否將第一行視為是資料的名稱
append=FALSE	若有相同名稱的檔案, 是否將資料加在原檔案之後
quote=" "	引號的字元,若沒有引號以quote=" "表之
row.names	列的名稱
col.names	行的名稱
na.strings=" . "	" . "代表NA(遺失值,missing value)
nrows	列數

附表A.4b write.table的其他指令

指令	說明
header=TRUE	讀取時, 是否將第一行視為是資料的名稱
quote=TRUE	是否包含引號,若設定為TRUE,則行的字元(character)或因子(factor)會以" "(雙引號)括住
sep=","	資料以" , "作分隔
row.names=TRUE	是否輸出列的名稱
col.names=TRUE	是否輸出行的名稱
na="NA"	遺失值以NA表之

附表A.5 圖形

指令	說明
barplot()	長條圖
boxplot()	盒鬚圖
hist()	直方圖
pairs()	散佈圖矩陣
pie()	圓餅圖
plot()	散佈圖
qqnorm()	常態分佈之Q-Q圖
stem()	莖葉圖
ts.plot()	時間序列圖

附表A.5a 其他圖形指令

指令	說明
abline()	將直線(給定截距及斜率)加至圖形之中
arrows()	畫出點與點之間的箭頭
axis()	軸
col	色彩的設定,可以red,blue等顏色指令,或灰階(gray20),或rainbow等設定圖形的色彩
legend()	在圖形中增加圖例
lines()	將線加至圖形之中
lty	線段的樣式
par(mfrow=c(1,2))	將兩張圖畫在同一視窗中,根據列作分割
par(mfcol=c(1,2))	將兩張圖畫在同一視窗中,根據行作分割
pch	點的樣式
points()	將點加至圖形之中
segments()	畫出點與點之間的線段
text()	在圖形中增加文字
title()	圖形的標題
x11()	新增畫圖視窗
xlim,ylim	x軸和y軸的設定

附表A.6 數學上的函數

指令	說明
log(x)	對x取自然對數
log10(x)	對x取對數(以10為底)
exp(x)	e^x
sin(x)	取正弦值
cos(x)	取餘弦值
tan(x)	取正切值
asin(x)	取反正弦值
acos(x)	取反餘弦值
atan(x)	取反正切值
min(x)	取x中的最小值
max(x)	取x中的最大值

附表A.7 統計上的函數

指令	說明
cov(x,y)	x,y的共變異數
cor(x)	x的相關係數
IQR(x)	x的內四分位數間距
length(x)	x中元素個數的長度
mean(x)	x的平均值
median(x)	x的中位數
range(x)	x的全距
quantile(x)	x的第0,25,50,75,100分位數
sd(x)	x的標準差
sum(x)	x的和
var(x)	x的變異數

附表A.8 常態分佈

指令	說明
dnorm(x)	常態分佈中x所對應的密度
pnorm(q)	常態分佈中q所對應的累積機率
qnorm(p)	常態分佈中機率值為p所對應的分位點
rnorm(n)	利用常態分佈產生n個隨機亂數

附表A.8a 其他分佈

指令	說明
pnorm(x,mean,sd)	常態分佈
plnorm(x,mean,sd)	對數常態分佈
pt(x,df)	$\boldsymbol{t}$分佈
pf(x,df1,df2)	F分佈
pchisq(x,df)	卡方分佈
pbinom(x,n,p)	二項分佈
ppois(x,lambda)	波松分佈
punif(x,min,max)	均勻分佈
pexp(x,rate)	指數分佈
pgamma(x,shape,scale)	Gamma分佈
pbeta(x,a,b)	Beta分佈

註:可如同常態分佈的指令,求得密度(d),分位點(q)及隨機亂數(r)。

附表A.9 統計方法

指令	說明
t.test	單或雙樣本之t檢定
pairwise.t.test	成對t檢定
var.test	變異數檢定或F檢定
chisq.test	卡方檢定
binom.test	二項檢定
fisher.test	費雪精準檢定
prop.test	二比例檢定
wilcox.test	單或雙樣本之Wilcoxon檢定
ks.test	K-S檢定
lm(y~x)	配適線性模型

附表A.10 函數及遞迴指令

指令	說明
function(變數列表){表示式}	自行設定函數
if(判斷式) 表示式	若判斷式為眞,則執行表示式之指令
ifelse(判斷式,眞,否)	若判斷式為眞,則執行"眞"的指令; 若為否,則執行"否"的指令
if(判斷式) 眞 else 否	指令同ifelse
for(變數 in seq) 表示式	for的迴圈設定,當變數滿足seq的範圍,則執行表示式之指令
while(判斷式) 表示式	while的迴圈設定,當while的判斷式為眞,則執行表示式之指令

附表A.10a 函數及遞迴指令的例子

例子	指令
f(x,y)=$x^2+\sqrt{y}$	function(x,y) { x^2+sqrt(y) }

若i<5,則列出"TRUE";
若i$\not<$5,則列出"FALSE";
	f<−function(i) {
	if(i<5)
	print("TRUE")
	else
	print("FALSE") }

for的迴圈,列出i的值	for(i in 1:3) print(i)
while的迴圈,列出i的值	i<−1 while(i<=3) { print(i) i<−i+1 }

附　表 B

附表B.1 台北、高雄及台中之月平均溫度數據資料

月份	一月	二月	三月	四月	五月	六月
台北	15.8	15.9	18	21.7	24.7	27.4
高雄	18.8	19.7	22.3	25.2	27.2	28.4
台中	16.2	16.8	19.4	23	25.7	27.5

月份	七月	八月	九月	十月	十一月	十二月
台北	29.2	28.8	27.1	24.3	20.9	17.6
高雄	28.9	28.3	27.9	26.4	23.4	20.2
台中	28.5	28	27.2	24.9	21.4	17.8

附表B.2 92年翠水庫日平均水位數據資料

	一月	二月	三月	四月	五月	六月
1	155	154.17	147.64	141.03	142.4	136.41
2	155.23	153.97	147.33	140.85	142.23	136.05
3	155.43	153.81	147.06	140.71	142.07	135.67
4	155.53	153.65	146.78	141.25	141.87	135.27
5	155.55	153.47	146.51	141.89	141.69	134.94
6	155.55	153.29	146.25	142.31	141.63	134.57
7	155.63	153.06	146	142.57	142.06	134.36
8	156.07	152.83	145.73	142.66	142.09	134.85
9	156.44	152.63	145.52	142.69	141.99	135.24
10	156.75	152.4	145.34	142.93	141.83	135.38
11	156.98	152.17	145.18	143.21	141.64	135.36
12	157.1	151.93	145	143.44	141.41	135.39
13	157.15	151.71	144.8	143.49	141.21	135.42
14	157.15	151.51	144.6	143.54	140.97	135.72
15	157.11	151.28	144.35	143.57	140.73	137.98
16	157.06	151.05	144.08	143.47	140.61	139.02
17	156.97	150.86	143.84	143.36	140.51	140.68
18	156.83	150.62	143.59	143.24	140.35	142.72
19	156.7	150.37	143.31	143.06	140.19	144.74
20	156.57	150.1	143.06	142.86	139.98	145.83
21	156.43	149.84	142.89	142.69	139.78	147.45
22	156.25	149.59	142.67	142.69	139.47	148.05
23	156.03	149.33	142.46	143.08	139.26	148.39
24	155.86	149.06	142.27	143.22	138.97	148.64
25	155.65	148.77	142.06	143.2	138.68	148.83
26	155.47	148.49	141.84	143.16	138.4	148.98
27	155.31	148.23	141.58	143.09	138.08	149.11
28	155.08	147.92	141.31	142.96	137.78	149.17
29	154.87		141.35	142.82	137.46	149.16
30	154.65		141.34	142.64	137.15	149.17
31	154.36		141.19		136.77	

附表B.2 92年翠水庫日平均水位數據資料(續)

	七月	八月	九月	十月	十一月	十二月
1	149.11	142.39	133.5	142.11	148.34	159.48
2	148.98	142.02	133.57	141.94	148.15	159.72
3	148.85	141.65	134.21	141.78	147.98	159.89
4	148.71	141.28	134.78	141.89	147.83	160.03
5	148.54	140.93	134.96	142.21	147.77	160.15
6	148.36	140.57	134.88	142.43	147.64	160.25
7	148.21	140.18	134.75	143.79	147.47	160.33
8	148.09	139.77	134.62	145.17	147.25	160.41
9	148.1	139.3	134.4	146.03	147.06	160.58
10	148.01	138.89	134.53	146.54	147.36	160.78
11	147.9	138.54	138.12	146.85	147.81	160.97
12	147.84	138.18	140.11	147.06	148.25	161.12
13	147.72	138.13	140.9	147.25	148.56	161.28
14	147.57	138.1	141.39	147.66	148.72	161.43
15	147.36	138.13	141.7	148.39	148.79	161.5
16	147.16	138.02	141.92	148.97	148.77	161.52
17	146.92	137.74	142.1	149.32	149.5	161.5
18	146.67	137.54	142.16	149.55	150.45	161.45
19	146.41	137.38	142.15	149.72	151.1	161.38
20	146.19	137.37	142.11	149.81	151.46	161.28
21	145.95	137.33	142.08	149.85	151.67	161.19
22	145.64	137.12	142.07	149.83	151.97	161.09
23	145.32	136.81	142.02	149.78	152.39	160.95
24	145.02	136.52	142	149.72	152.74	160.8
25	144.72	136.22	141.96	149.61	153.06	160.64
26	144.36	135.88	141.9	149.46	153.69	160.49
27	144.05	135.53	142.03	149.33	154.7	160.31
28	143.78	135.15	142.29	149.16	156.53	160.13
29	143.42	134.74	142.33	148.96	158.34	159.98
30	143.08	134.29	142.26	148.78	159.1	159.8
31	142.74	133.85		148.58		159.62

索 引

中 文 索 引

【二畫】

二項分佈, 47

二樣本之t分佈, 66

【三畫】

下四分位數, 26, 40

下管制限, 97

上四分位數, 26, 41

上管制限, 97

【四畫】

不合格率管制圖, 104

不良率管制圖, 104

中位數, 26, 40, 41

內四分位數間距, 41

分位圖, 29

分位數, 26, 40

太陽黑子, 30, 119

【五畫】

加拿大山貓, 116

卡方分佈, 50

卡方檢定, 69

平均值, 40

母體變異數之檢定, 68

皮爾生之卡方適合度檢定, 69

【六畫】

全距, 39

全距管制圖, 97

共變異數, 43

印刷電路板, 107

向量, 16

老忠實噴泉, 22, 87

自由度, 50

【七畫】

伯努力分佈, 47

伯努力試驗, 47

刪除, 12

決定係數, 89

【八畫】

孟德爾, 70

波松分佈, 49

直方圖, 23

長條圖, 23

長鬚圖, 25

【九畫】

信賴係數, 80

信賴區間, 80, 89

指數加權移動平均管制圖, 111

查詢, 12

柳橙汁, 105
活塞環, 99
相關係數, 43
計量, 97
計量管制圖, 97
計數值, 104
計數管制圖, 97

【十畫】
個人電腦製造商, 110
時間序列, 30, 115
時間序列圖, 30
矩陣, 16
缺點數管制圖, 104
陣列, 16

【十一畫】
偏斜性, 25
區間估計, 80
常態分佈, 51
常態分佈的Q-Q圖, 29
常態性, 29
清單, 13
盒鬚圖, 25
第p分位數, 40
第25分位數, 40
第50分位數, 40
第75分位數, 41
累積和管制圖, 111
莖葉圖, 22

【十二畫】
最大值, 26, 39
最大概似估計法, 77
最大概似估計值, 77
最小平方法, 85
最小值, 26, 39
單因子變異數分析方法, 90
單位缺點管制圖, 104
單樣本之t檢定, 63
幾何分佈, 48
散佈圖, 26
散佈圖矩陣, 28
無母數檢定, 70
程式庫, 18

【十三畫】
匯入檔案, 16
匯出檔案, 15
經驗分佈函數, 71
資料表, 12, 16

【十四畫】
管制圖, 97
翡翠水庫, 39
說明, 14
鳶尾, 28

【十五畫】

樣本變異數, 42
標準差, 42
標準差管制圖, 97
線上查詢, 14
適合度檢定, 69
隨機亂數, 29

【十八畫】

簡單線性迴歸, 85
簡單線性迴歸模型, 85
轉置, 18
離群值, 25

【二十一畫】

鐵達尼號, 24

【二十三畫】

變更路徑, 10
變異數分析, 90
變異數分析表, 88
變異數比例檢定, 68
變異數的檢定, 68
變數, 11

英文索引

(粗體表R的指令)

+ , 9

> , 9

<−, 11

R^2, 89

χ^2 test, 69

$\mathcal{F}$ distribution, 54

$\mathcal{F}$ 分佈, 54

$\boldsymbol{t}$ distribution, 53

$\boldsymbol{t}$ 分佈, 53

$\overline{X}$control chart, 97

$\overline{X}$管制圖, 97

π, 60

p-value, 65

p值, 65

【A】

abline, 93

ACF, 116

acf, 116, 124

add, 34

airmiles, 18

alternative, 73

analysis of variance , 88, 90

ANOVA, 90

aov, 88, 94

apply, 46

AR, 115

ar, 125

arima, 122, 126

ARIMA模型, 115

array, 16

as.matirx, 18, 33

at, 34

attribute, 104

attributes control charts, 97

autocorrelation function, 116

【B】

bar plot, 23

barplot, 23, 33

base, 18

Bernoulli trial, 47

beside, 33

beta distribution, 52

beta 分佈, 52

binomial distribution, 47

box-and-whisker plot, 25

boxplot, 25, 34

boxwex, 34

breaks, 33

byrow, 18

【C】

c, 16

c chart, 97

change dir, 10
chisq.test, 70, 74
chi-squared distribution, 50
coefficient of determination, 89
col, 32, 45
col.names, 16
conf.level, 73
confit, 89
control chart for fraction nonconforming,104
control chart for nonconformities, 104
control chart fot nonconformities per unit, 104
cor, 43
correlation, 43
correlation coefficient, 43
cov, 43
covariance, 43
cumulative-sum control chart, 111
cusum, 111, 114
c管制圖, 97

【D】
data frame, 12, 16
dbinorm, 58
degrees of freedom, 50
diff, 122, 125
dim, 17

【E】
empirical distribution function, 71
esponentially weighted moving-average control chart, 111
EWMA, 111
ewma, 111, 114
expression, 60

【F】
factor, 95
Feitsui Reservoir, 39
fitdistr, 78
freq, 32
function, 58
F-檢定, 68

【G】
gamma distribution, 50
Gamma 分佈, 50
geometric distribution, 48
goodness of fit, 69

【H】
header, 16
help, 14, 19
help.search, 14
help.start, 14
hist, 23, 32
histogram, 23

horiz, 33

【I】

if, 95
interquantile range , 41
interval estimation, 80
IQR, 41
iris, 28
is.factor, 95
is.matrix, 18, 33

【K】

Kolmogorov-Smirnov検定, 70
ks.test, 71, 74
K-S検定, 70

【L】

labels, 32
LCL, 97
legend, 33, 58
list, 13
lm, 87, 93
low control limit, 97
ls, 12
ls.print, 93
ls.str, 12
lsfit, 86
lynx, 116

【M】

MA, 115
main, 35
matlines, 60
matplot, 58
matrix, 16, 17
max, 39
maximum likelihood estimate, 77
maximum value, 39
mean, 40, 41
median, 40
method of maximum likelihood, 77
mfcol, 32
mfrow, 32
min, 39
minimum value, 39
MLE, 77

【N】

na.strings, 16
normal distribution, 51
Normal Q-Q plot, 29
Normal Quantile-Quantile plot, 29
normality, 29
nrow, 18

【O】

objects, 12
Old Faithful geyser, 22

one-way ANOVA, 91
online help, 14
orange juice, 105
outer, 58
outliers, 25

【P】
P chart, 97
PACF, 116
pacf, 116, 124
package, 18
paired, 74
pairs, 28, 36
par, 32
partial autocorrelation function, 116
pch, 36
pcmanufact, 110
Pearson's chi-squared goodness-of-fit test, 69
piston rings , 99
plot, 30
plot.ecdf, 75
plot.stepfun, 74
Poisson distribution, 49
pth quantile, 40
P管制圖, 97

【Q】
qcc, 97, 113
Q-Q plot, 29
qqline, 37
qqnorm, 29, 37
quantile, 41
Quantile-Quantile plot, 29
quote, 16

【R】
R chart, 97
rainbow, 45
random number, 29
range, 39
ratio, 74
read.table, 16
rm, 12
row.names, 16
rt, 37
runif, 74
R管制圖, 97

【S】
S chart, 97
sample variance, 42
scatter plot, 26
scatterplot matrices, 28
sd, 42
sep, 16

skewness, 25
S-PLUS, 1
standard deviation, 42
stem, 22, 32
stem-and-leaf diagram, 22
summary, 93
sunspots, 30, 119
S管制圖, 97

【T】
t, 18
t.test, 65, 73, 81
t.test$conf.int, 81
text, 93
time series, 30, 115
time series plots, 30
Titanic, 24
ToothGrowth, 25
ts, 116
ts.plot, 30
type, 36

【U】
u chart, 97
UCL, 97
upper control limit, 97
u管制圖, 97

【V】
var, 42
var.equal, 74
var.test, 68, 74
variable, 11, 97
variable control charts, 97
variance, 42
vector, 16

【W】
write.table, 15

【X】
xlab, 35
xlim, 32

【Y】
ylab, 35
ylim, 33