

# R 軟體：基本統計分析快速入門

作者：淡江大學統計系 陳景祥 2010/03/15

## 前言：

這份文件是給尚未學習 R 軟體，但想要快速了解 R 軟體在一般統計分析操作程序的讀者一個初步的概念

假設我們已經安裝了 R 軟體，也已經進入 R 軟體的視窗界面

假設我們要分析一組學生的資料，資料儲存於 c:\mydir\students.txt，數字之間以一個或多個空格分開，內容如下：

```
studentID gender bloodtype age score
U101      M      AB  18   43
U102      M      B   19   56
U103      M      AB  17   52
.....
U148      F      AB  21   58
U149      F      B   22   57
U150      M      O   17   75
```

這個範例資料檔可於以下網址下載：

<http://steve-chen.net/R/students.txt>

## 1. 讀入資料

```
> students = read.table("c:/mydir/students.txt", header = TRUE )
> students
```

```
  studentID gender bloodtype age score
1      U101      M      AB  18   43
2      U102      M      B   19   56
3      U103      M      AB  17   52
.....
```

## 說明：

- (1). read.table 可以讀入外部文字檔，並儲存為 data.frame 變數 students
- (2). read.table 函數裡面的 header = TRUE 表示資料檔第一橫列是變數名稱
- (3). Windows 作業系統中的檔案路徑斜線需用反斜線 "/"

## 2. 查詢資料：使用 \$ 符號查詢 students 中的各個變數

查詢 score 變數：

```
> students$score
[1] 43 56 52 54 55 71 66 31 61 67 57 54 56 56 54 59 39 58 55 42 53 67 73 45
[25] 49 56 69 44 68 49 49 57 69 51 47 51 45 46 51 73 61 52 60 48 43 49 62 58
[49] 57 75
> students$score[3]           # 查詢第 3 個學生的成績
[1] 52
```

查詢 bloodtype 變數：

```
> students$bloodtype
[1] A B B A B O A B O O A B B A A A B A B B A A A B O A B B O
[25] B A B B A O A B O B A B O O O O A B A A B A B B O A B A A B A B
[49] B O
Levels: A A B B O
```

```
> students$bloodtype[13]      # 查詢第 13 個學生的血型
[1] A
Levels: A A B B O
```

說明：

1. R 軟體自動將文字變數轉成 Factor 分類變數型態，Levels A A B B O 表示這組變數共有四個分類
2. 若  $x$  是一個向量變數 (vector)，則  $x[k]$  是  $x$  第  $k$  個元素的值

## 3. 基本敘述統計

```
> summary(students$score)           # 最小值, Q1, 中位數, Q3, 最大值
  Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
 31.00  49.00   55.00   55.26  60.75   75.00
> mean(students$score)              # 平均數
[1] 55.26
> sd(students$score)                # 標準差 (standard deviation)
[1] 9.569957
```

### 說明：

其他常用的敘述統計函數還包含 sum(總和), range (全距), var (樣本變異數), length (樣本數/向量長度), fivenum, mad, IQR, median, min, max 等。

```
> tapply(students$score, students$gender, mean) # 男女兩分類的平均成績
```

```
      F      M  
55.39130 55.14815
```

```
> tapply(students$score, students$bloodtype, mean) # 四種血型的平均成績
```

```
      A      AB      B      O  
57.50000 52.31250 57.30769 55.46154
```

性別與血型組合下的八種分類之平均成績：

```
> tapply(students$score, list(students$gender, students$bloodtype), mean)
```

```
      A  AB  B  O  
F 62.50000 56.5 56.00 51.00000  
M 55.83333 49.8 60.25 59.28571
```

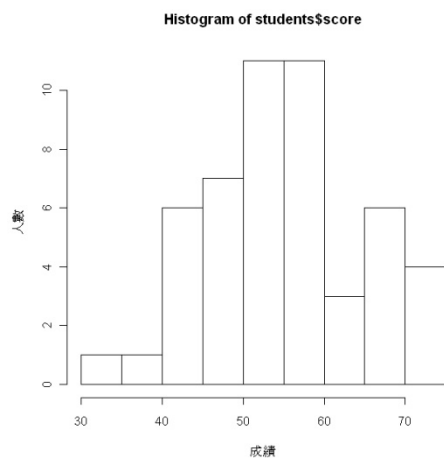
說明：tapply 函數是將某個計算函數套用到某些分類。其語法是

tapply(數值向量, 分類向量, 彙整函數的名稱)

或 tapply(數值向量, list(分類向量 1, 分類向量 2, ...), 彙整函數的名稱)

## 4. 基本統計圖形

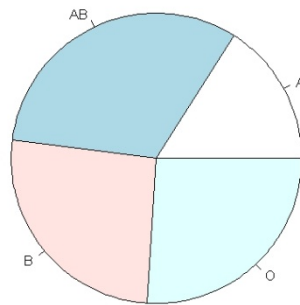
```
> hist(students$score, xlab="成績", ylab="人數") # 直方圖(histogram)
```



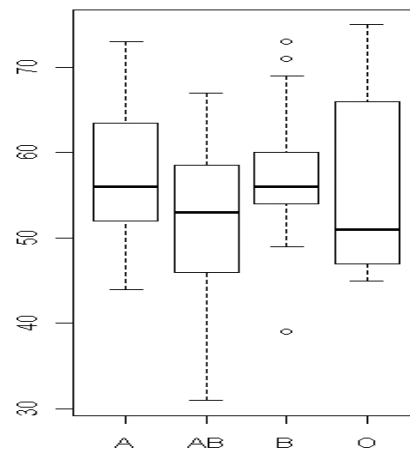
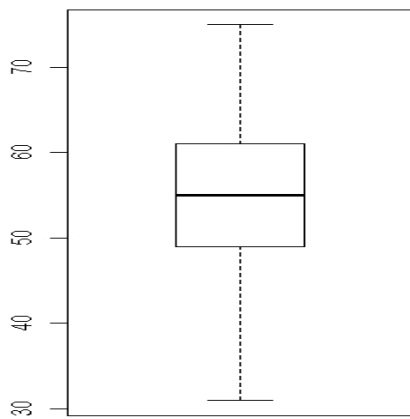
```
> table(students$bloodtype) # 計算四種血型的人數
```

A AB B O  
8 16 13 13

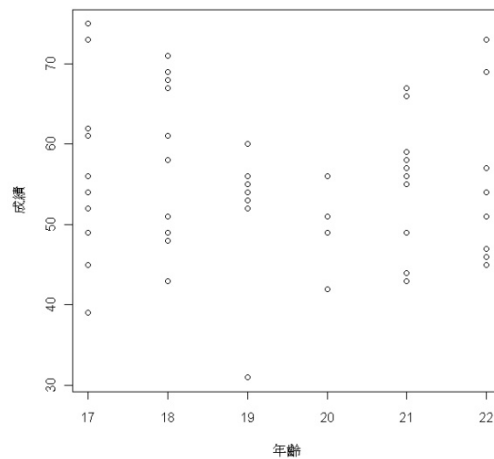
```
> pie(table(students$bloodtype)) # 血型人數的圓餅圖
```



```
> boxplot(students$score)  
> boxplot(students$score ~ students$bloodtype) # 四種血型下的成績盒鬚圖
```



```
> plot(students$age, students$score, xlab="年齡", ylab="成績") # X-Y 散佈圖
```



## 5. 基本統計檢定

(1) 母體平均數檢定：平均成績是否等於 60 分？

```
> t.test(students$score,mu=60,alternative="two.sided")
```

One Sample t-test

data: students\$score

t = -3.5023, df = 49, p-value = 0.0009945

alternative hypothesis: true mean is not equal to 60

95 percent confidence interval:

52.54025 57.97975

sample estimates:

mean of x

55.26

說明：

t.test(x, mu = H0 中假設的  $\mu$  值, alternative = "two.sided"、"greater"、或 "less")

(2) 母體比例檢定：男性學生的比例是否等於 0.5 ？

```
> table(students$gender) # 計算男性與女性的人數
```

F M

23 27

也可使用 sum 函數計算男性學生人數：

```
> sum(students$gender == "M")
```

[1] 27

```
> length(students$gender) # 計算總人數
```

```
[1] 50
```

```
> prop.test(27,50,p=0.5)
```

```
1-sample proportions test with continuity correction
```

```
data: 27 out of 50, null probability 0.5
```

```
X-squared = 0.18, df = 1, p-value = 0.6714
```

```
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.3945281 0.6793659
```

```
sample estimates:
```

```
p
```

```
0.54
```

說明：prop.test(x, n, p=H<sub>0</sub> 中假設的 p 值)

(3) 卡方檢定：性別與血型是否互相獨立？

```
> tab = table(students$gender,students$bloodtype) # 用 table 函數計算人數
```

```
> tab
```

```
  A AB  B  O
```

```
F  2  6  9  6
```

```
M  6 10  4  7
```

```
> chisq.test(tab)
```

```
Pearson's Chi-squared test
```

```
data: table(students$gender, students$bloodtype)
```

```
X-squared = 4.7101, df = 3, p-value = 0.1943
```

## 6. 相關分析與迴歸分析

(1) 年齡與成績的樣本相關係數

```
> cor(students$age, students$score)
```

```
[1] -0.0862235
```

(2) 簡單線性迴歸分析：應變數 score，解釋變數 age

```
> result1 = lm(score ~ age, data=students) # lm : linear model
```

```
> result1
```

```
Call:
```

```
lm(formula = score ~ age, data = students)
```

Coefficients:

```
(Intercept)      age
      63.9615    -0.4481
```

```
> summary(result1)
```

Call:

```
lm(formula = score ~ age, data = students)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-24.4482  -6.6722  -0.2242   4.6297  18.8960
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.9615    14.5758   4.388 6.24e-05 ***
age          -0.4481     0.7473  -0.600  0.552
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.633 on 48 degrees of freedom

Multiple R-squared: 0.007434, Adjusted R-squared: -0.01324

F-statistic: 0.3595 on 1 and 48 DF, p-value: 0.5516

(3) 複迴歸分析：gender 與 bloodtype 預設當作分類變數處理

```
> result2 = lm(score ~ age + gender + bloodtype, data=students)
```

```
> result2
```

Call:

```
lm(formula = score ~ age + gender + bloodtype, data = students)
```

Coefficients:

```
(Intercept)      age      genderM  bloodtypeAB  bloodtypeB
      68.8864    -0.5676    -0.2355     -5.5362     -0.5039
bloodtypeO
      -2.2084
```

```
> summary(result2)
```

Call:

```
lm(formula = score ~ age + gender + bloodtype, data = students)
```

Residuals:

Min	1Q	Median	3Q	Max
-21.331	-7.158	-1.414	7.319	18.206

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	68.8864	16.6205	4.145	0.000152 ***
age	-0.5676	0.7873	-0.721	0.474766
genderM	-0.2355	3.0082	-0.078	0.937944
bloodtypeAB	-5.5362	4.2813	-1.293	0.202723
bloodtypeB	-0.5039	4.6185	-0.109	0.913619
bloodtypeO	-2.2084	4.4466	-0.497	0.621918

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.773 on 44 degrees of freedom

Multiple R-squared: 0.06348, Adjusted R-squared: -0.04295

F-statistic: 0.5965 on 5 and 44 DF, p-value: 0.7027

## 7. 變異數分析(ANOVA)

(1) 一因子設計(Oneway Design)：應變數 score, 因子爲 bloodtype

```
> aov.result1 = aov(score ~ bloodtype, data=students)
> aov.result1
```

Call:

```
aov(formula = score ~ bloodtype, data = students)
```

Terms:

	bloodtype	Residuals
Sum of Squares	234.183	4253.438
Deg. of Freedom	3	46

Residual standard error: 9.615926  
Estimated effects may be unbalanced

一因子設計的 ANOVA 表格

```
> anova(aov.result1)
```

Call:

```
aov(formula = score ~ bloodtype, data = students)
```

Terms:

	bloodtype	Residuals
Sum of Squares	234.183	4253.438
Deg. of Freedom	3	46

Residual standard error: 9.615926  
 Estimated effects may be unbalanced

(2) 二因子設計(Twoway Design)：應變數 score, 因子為 bloodtype 與 gender

```
> aov.result2 = aov(score ~ bloodtype + gender, data=students)
> aov.result2
```

Call:

```
aov(formula = score ~ bloodtype + gender, data = students)
```

Terms:

	bloodtype	gender	Residuals
Sum of Squares	234.183	1.036	4252.402
Deg. of Freedom	3	1	45

Residual standard error: 9.720999  
 Estimated effects may be unbalanced

二因子設計的 ANOVA 表格

```
> anova(aov.result2)
```

Analysis of Variance Table

Response: score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bloodtype	3	234.2	78.061	0.8261	0.4865
gender	1	1.0	1.036	0.0110	0.9171
Residuals	45	4252.4	94.498		