Chapter 7

BASIC DATA ANALYSIS (2)

資料分析

當我們已經可以把資料放入變數後，接下來就是要進行資料分析的動作

這個章節將介紹利用簡單的內建函數以及繪圖函數

來解析資料的維度以及簡單的統計分佈

內建資料集 IRIS

IRIS 已經內建在 R 語言中 使用內建函數得到維度資料	
nrow	幾筆資料
ncol	幾個維度
dim	結合 nrow 和 ncol

```
1  iris_data <- iris
2
3  nr <- nrow(iris_data)
4  nc <- ncol(iris_data)
5  nd <-  dim(iris_data)
6
7  print(nr)
8  print(nc)
9  print(nd)
10
11
```

Console Terminal

F:/Course/1062/data mining/code/R/ch6/

```
> iris_data <- iris
>
> nr <- nrow(iris_data)
> nc <- ncol(iris_data)
> nd <-  dim(iris_data)
>
> print(nr)
[1] 150
> print(nc)
[1] 5
> print(nd)
[1] 150   5
>
```

內建資料集 IRIS

得到 IRIS 的更完整描述資料	
summary	描述統計
str	完整資料結構

```
1  iris_data <- iris
2
3  a <- summary(iris_data)
4  print(a)
5
6  str(iris_data)
7
```

```
> a <- summary(iris_data)
> print(a)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
> str(iris_data)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

hist() 直方圖

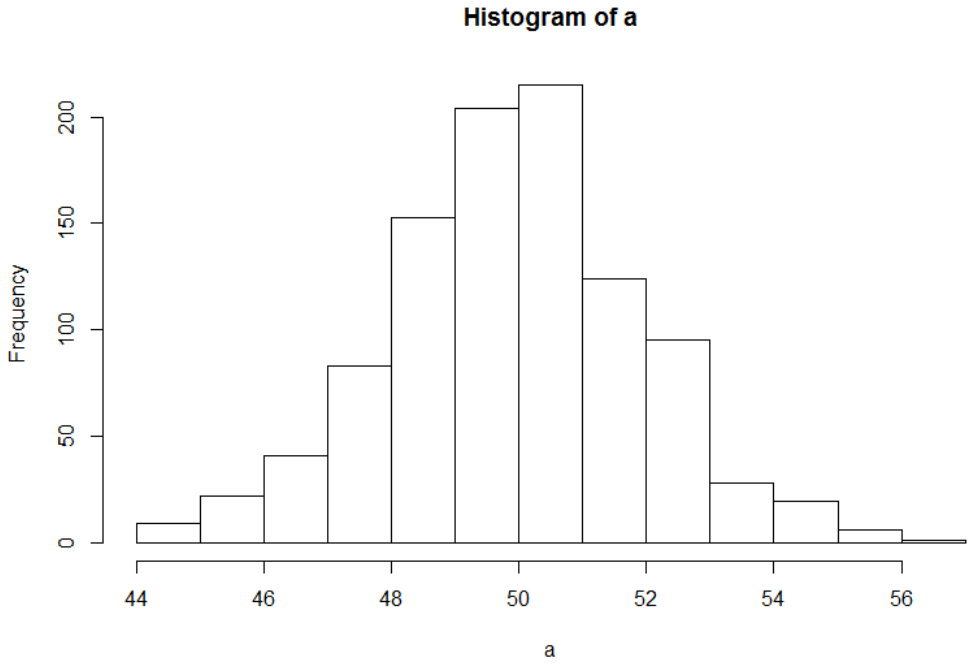
rnorm(data, 平均值 , 標準差)

使用 rnorm 產生標準常態分佈

平均值預設為 0

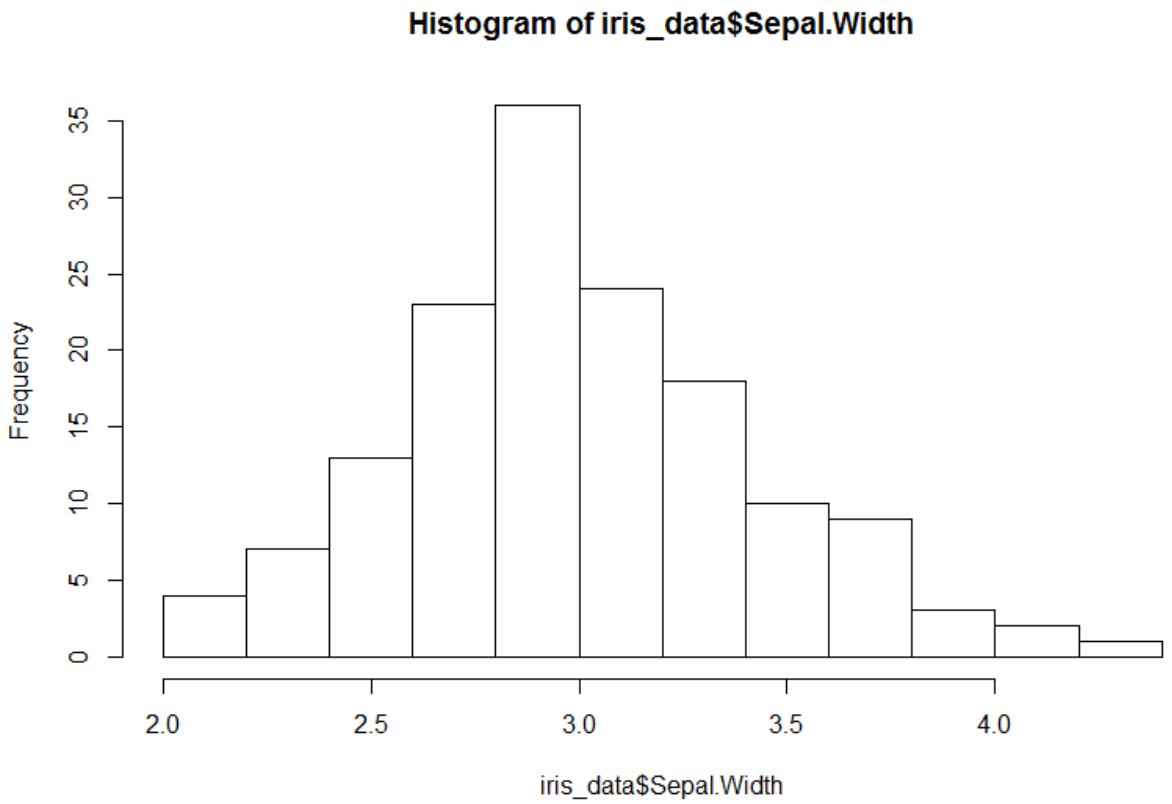
標準差預設為 1

```
1
2  a<- rnorm(1000,50,2)
3  hist(a)
4
```



hist() 直方圖

```
1  iris_data <- iris
2
3  hist(iris_data$Sepal.Length)
```

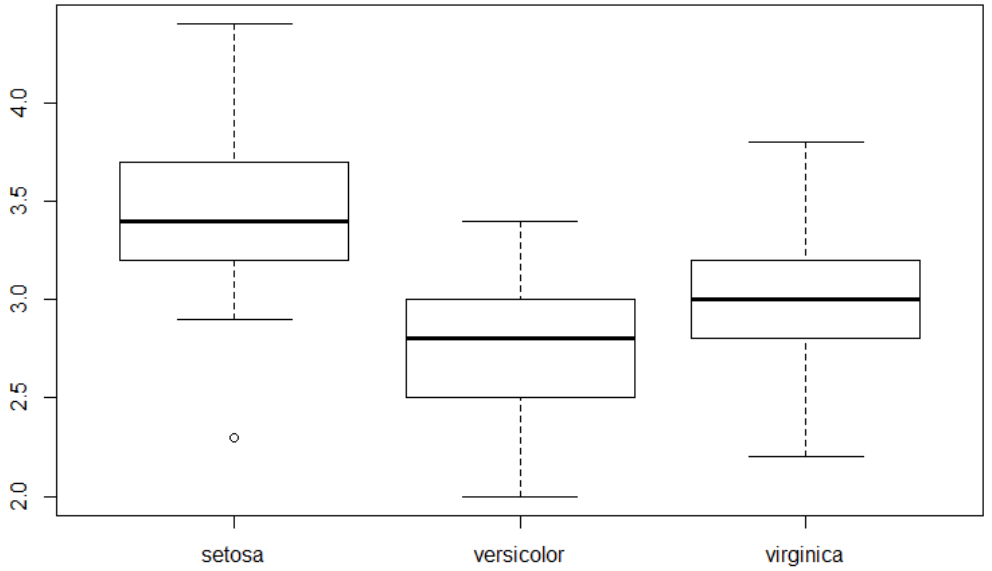


boxplot() 盒鬚圖

boxplot(維度 1~ 維度 2, data=“ 資料 ”)

可找出資料中
在維度 2 資料下找出
維度 1 的資料分佈

```
1  iris_data <- iris
2
3  boxplot(iris_data$Sepal.Width ~ iris_data$Species,
4          data= iris_data)
```



plot() 線圖

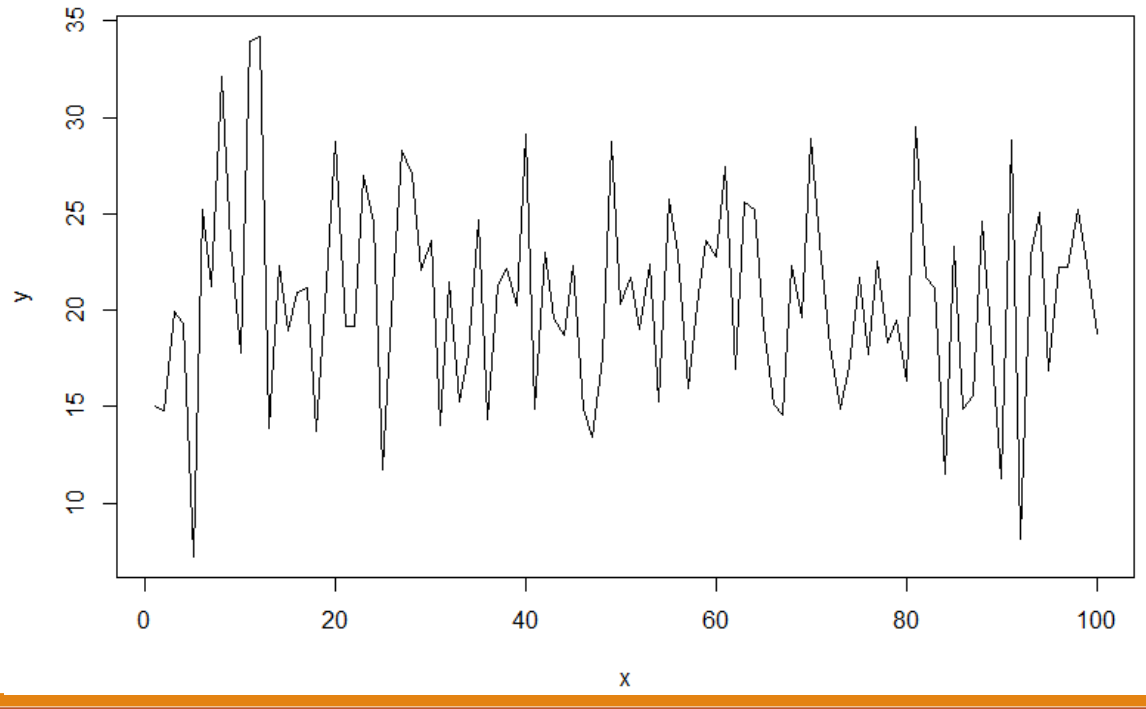
plot(data1, data2, … type =“l”)

可使用二筆資料

再搭配 type=“l” (L) 來畫出

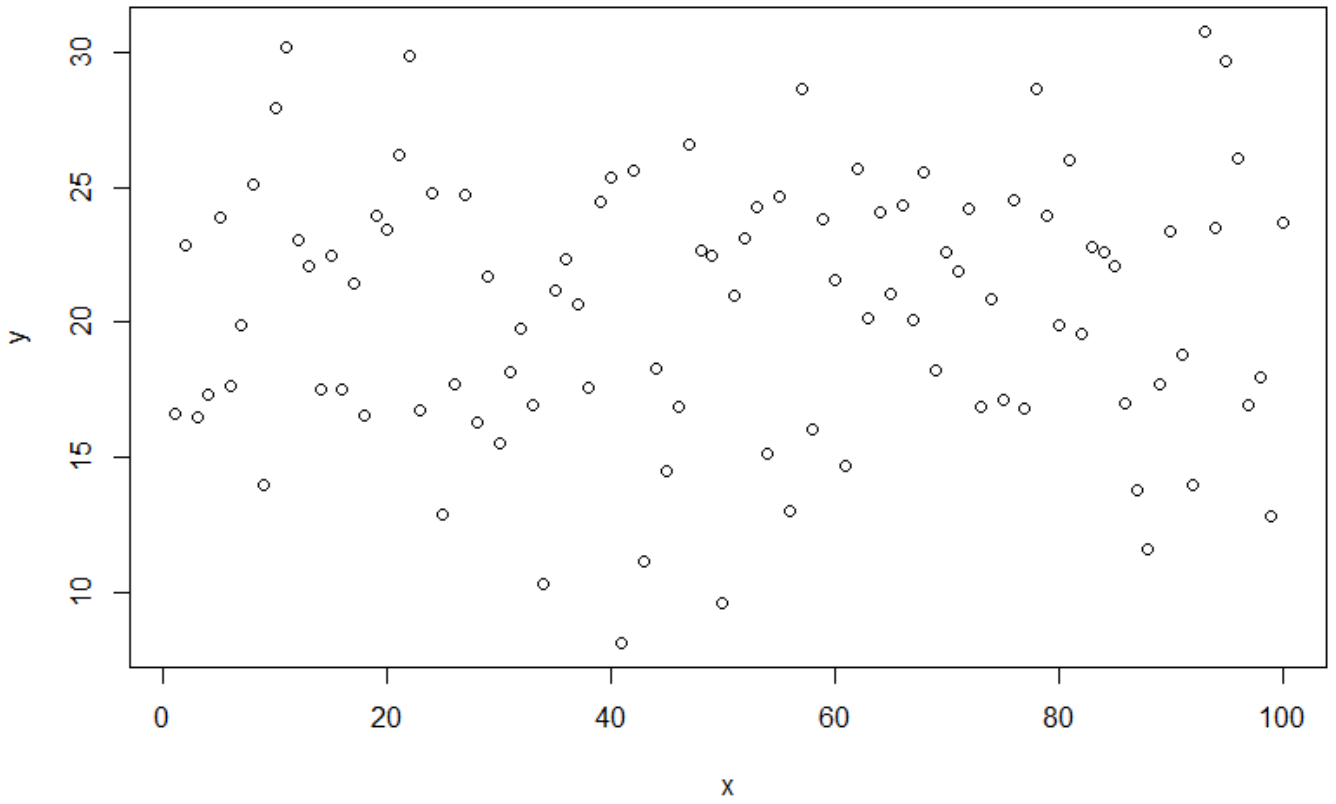
二維折線圖

```
1  #x 為 1->100
2  x <- seq(1:100)
3  #y 為亂數
4  y <- rnorm(100,20,5)
5  #畫出折線圖
6  plot(x, y, type = "l")
```



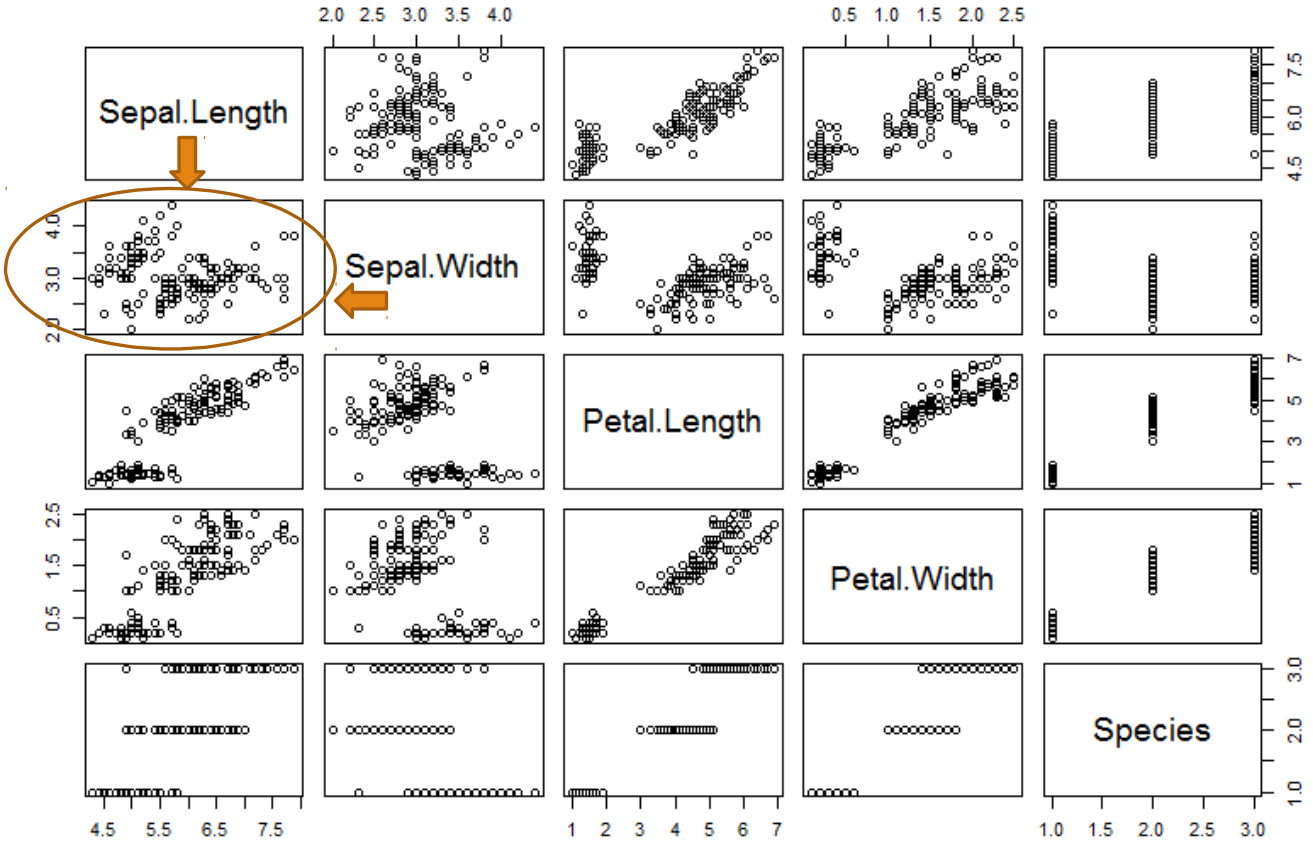
plot() 散佈圖

```
1  #x 為 1->100
2  x <- seq(1:100)
3  #y 為亂數
4  y <- rnorm(100,20,5)
5  #畫出散佈圖
6  plot(x, y)
```



plot() 散佈圖矩陣

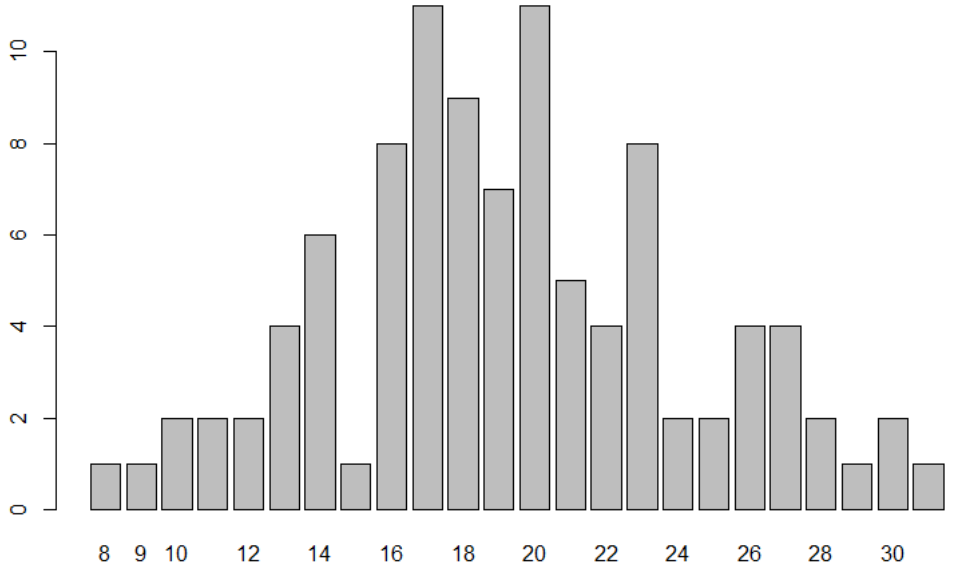
```
1  iris_data <- iris
2
3  plot(iris_data)
```



barplot() 長條圖

```r
x <- rnorm(100,20,5)
print(x)
#轉整數
int_x <- as.integer(x)
print(int_x)
#轉表格
tb_x <- table(int_x)
print(tb_x)
#畫出長條圖
barplot(table(int_x))
```

```
> x <- rnorm(100,20,5)
> print(x)
  [1] 17.584325 17.343265 17.061576 17.941511 23.545928 21.281984 10.718197
 [11]  8.462628 15.917764 20.137806 27.308930  9.935656 13.722779 14.598466
 [21] 11.938357 16.847238 17.197570 18.987094 28.114426 16.616147 20.381322
 [31] 14.746857 33.676046 20.462815 20.301268 19.667274 29.218228 23.319636
 [41] 16.075053 26.074742 19.055092 16.214007 23.960297 26.725900 16.527343
 [51] 22.031832 28.570993 19.698067 18.596489 22.427072 19.753277 23.138825
 [61] 20.069518 17.418920 14.047287 17.934654 21.859975 20.461715 23.467419
 [71] 13.359219 18.598325 25.848277 19.392618 11.813542 22.456916 21.409097
 [81] 26.495980 17.718516 20.053324 12.729554 16.361233 30.041202 27.490048
 [91] 30.630259 20.573149 17.030255 25.390337 14.502071 23.632821 27.204352
>
> int_x <- as.integer(x)
> print(int_x)
  [1] 17 17 17 17 23 21 10 10 19 20  8 15 20 27  9 13 14 20 21 18 11 16 17
 [37] 23 18 14 14 16 26 19 16 23 26 16 17 21 19 22 28 19 18 22 19 23 18 22
 [73] 25 19 11 22 21 17 20 26 26 17 20 12 16 30 27 18 16 13 30 20 17 25 14
>
> tb_x <- table(int_x)
> print(tb_x)
int_x
 8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 33
 1  1  2  2  2  4  6  1  8 11  9  7 11  5  4  8  2  2  4  4  2  1  2  1
```

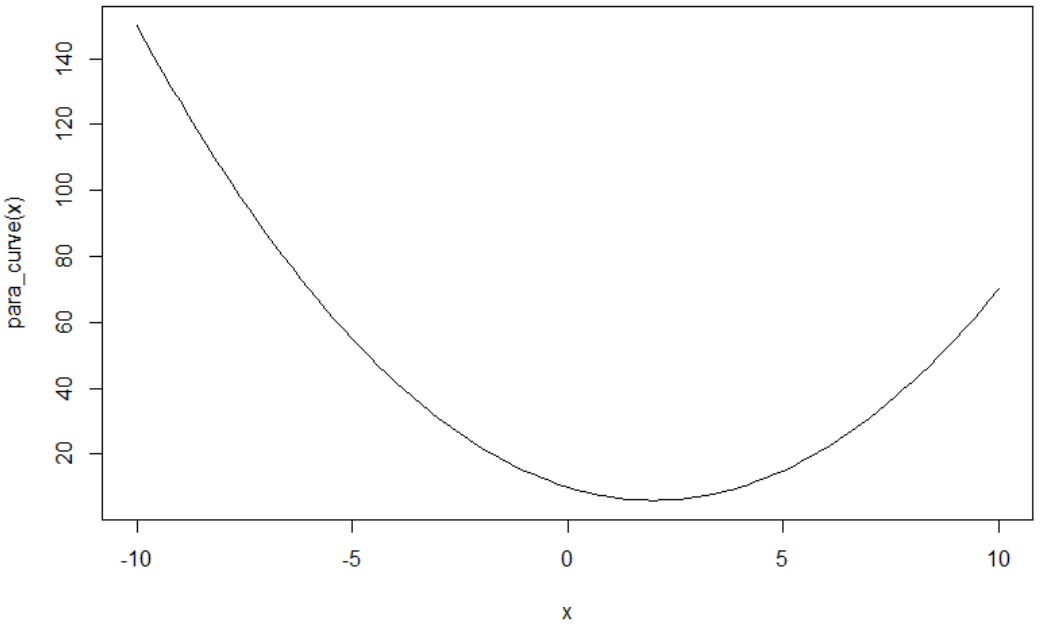


curve() 曲線圖

curve (data, from = “”, to =“”)

可利用 from – to
指定曲線圖起點和終點

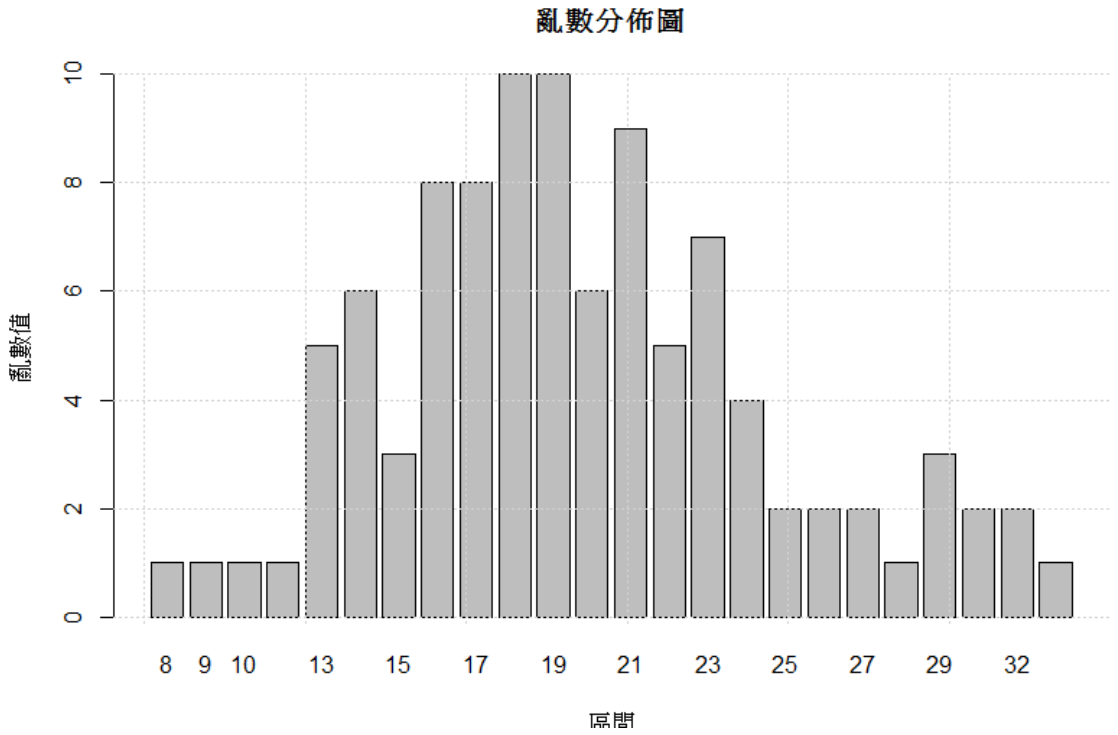
```
1  para_curve <- function(x)
2  {
3    return( (x-2)^2+6 )
4  }
5  curve(para_curve, from=-10, to=10)
```



自訂圖形元素

參數	說明
main	標題
xlab	X 軸標題
ylab	Y 軸標題
grid	加入格線

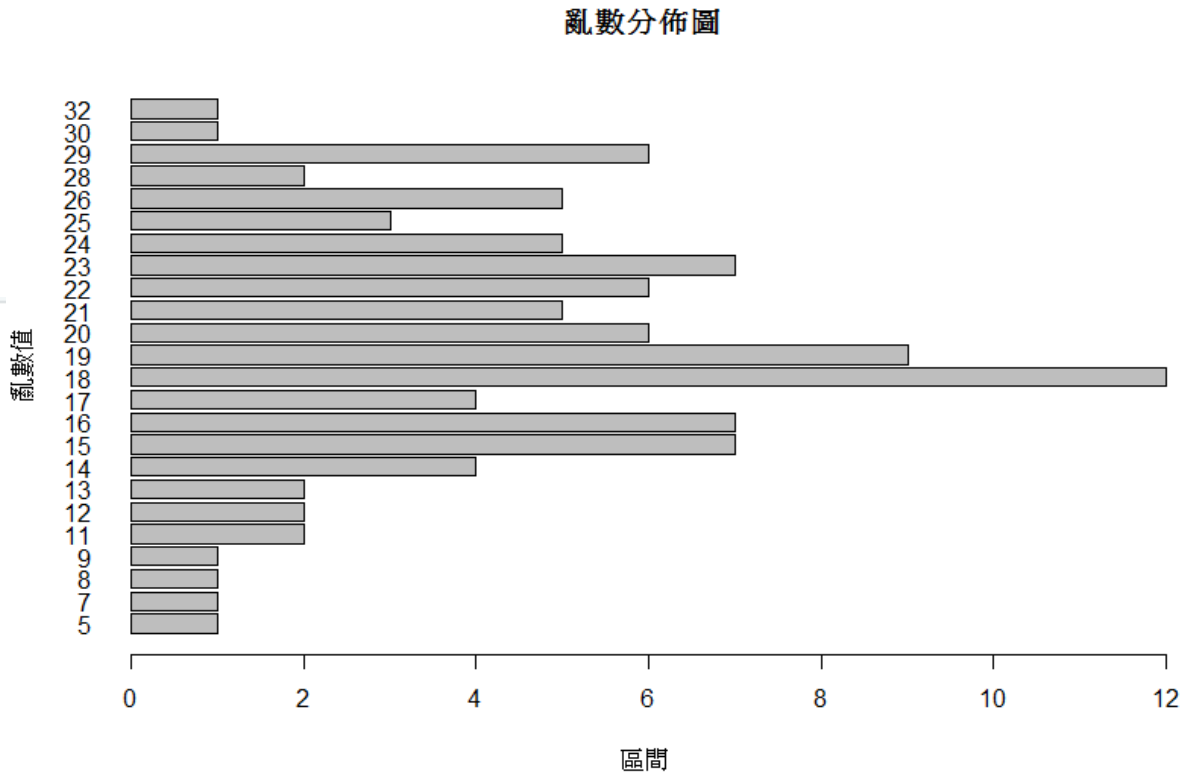
```
1  x <- rnorm(100,20,5)
2  int_x <- as.integer(x)
3  tb_x <- table(int_x)
4  barplot(table(int_x), main ="亂數分佈圖",
5          xlab = "區間",
6          ylab = "亂數值")
7  grid()
```



自訂圖形元素

參數	說明
horiz	水平
las	刻度顯示方向
cex.name	刻度大小 y 軸
cex.axis	刻度大小 x 軸

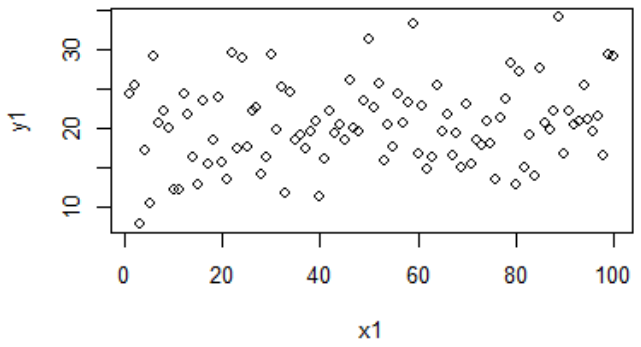
```
x <- rnorm(100,20,5)
int_x <- as.integer(x)
tb_x <- table(int_x)
barplot(table(int_x), main ="亂數分佈圖",
        xlab = "區間",
        ylab = "亂數值",
        horiz = TRUE,
        las=1,
        cex.name =1,
        cex.axis =1)
```

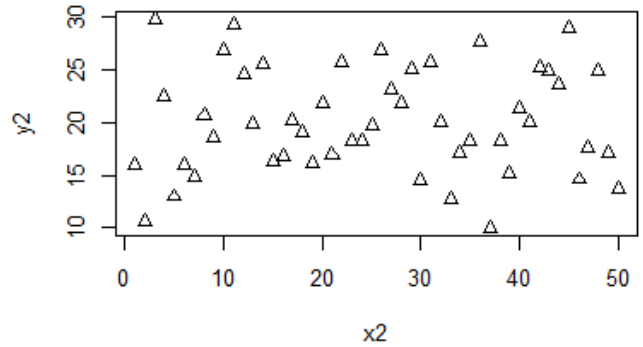


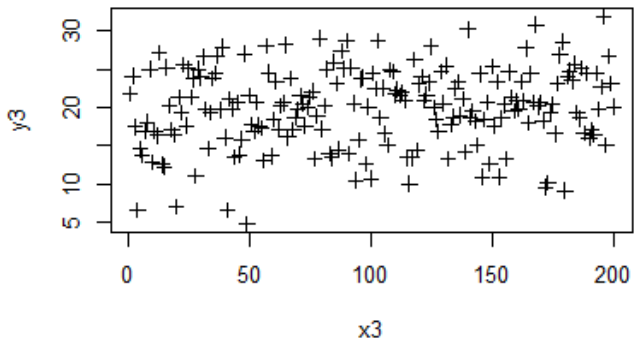
繪製多個圖形

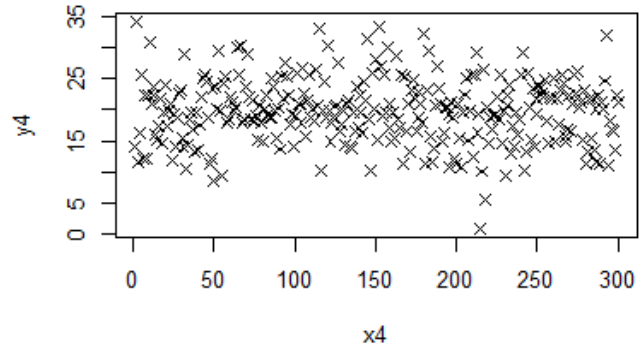
參數	說明
mfrow	切割區域 (m * n)

```
1  par(mfrow = c(2,2))
2  x1 <- seq(1:100)
3  y1 <- rnorm(100,20,5)
4
5  x2 <- seq(1:50)
6  y2 <- rnorm(50,20,5)
7
8  x3 <- seq(1:200)
9  y3 <- rnorm(200,20,5)
10
11 x4 <- seq(1:300)
12 y4 <- rnorm(300,20,5)
13
14 plot(x1, y1, pch =1)
15 plot(x2, y2, pch =2)
16 plot(x3, y3, pch =3)
17 plot(x4, y4, pch =4)
```

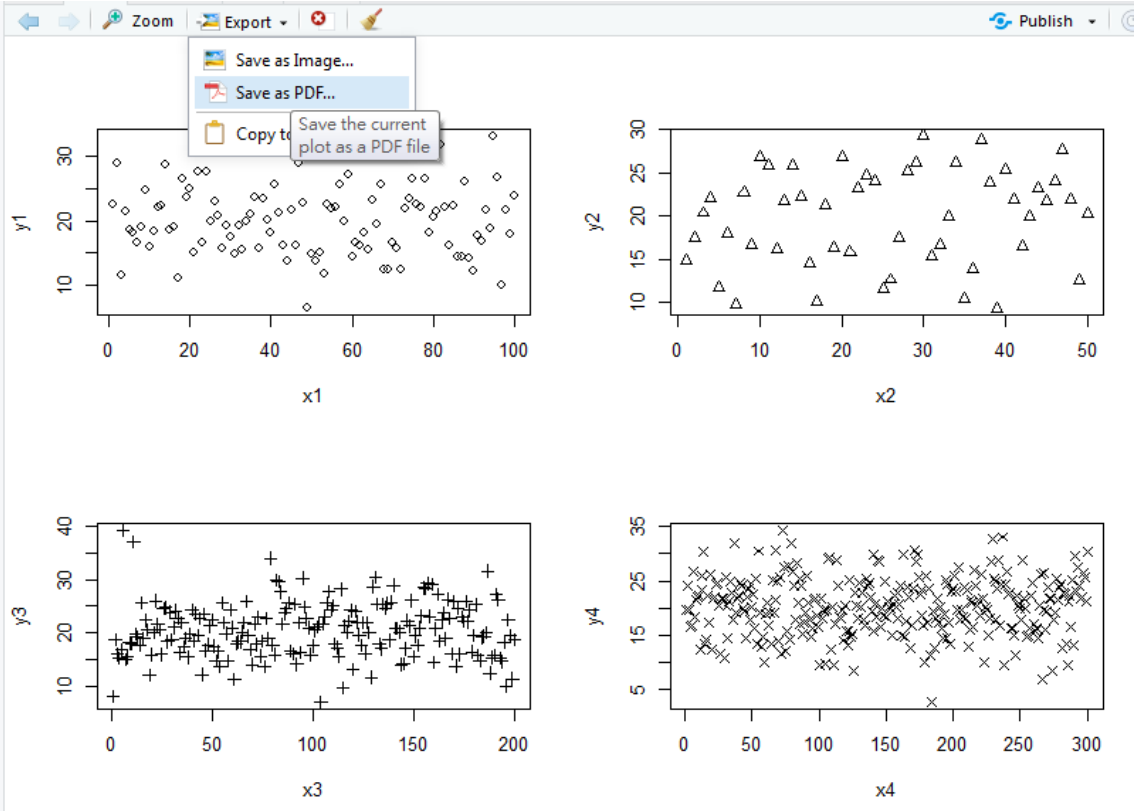


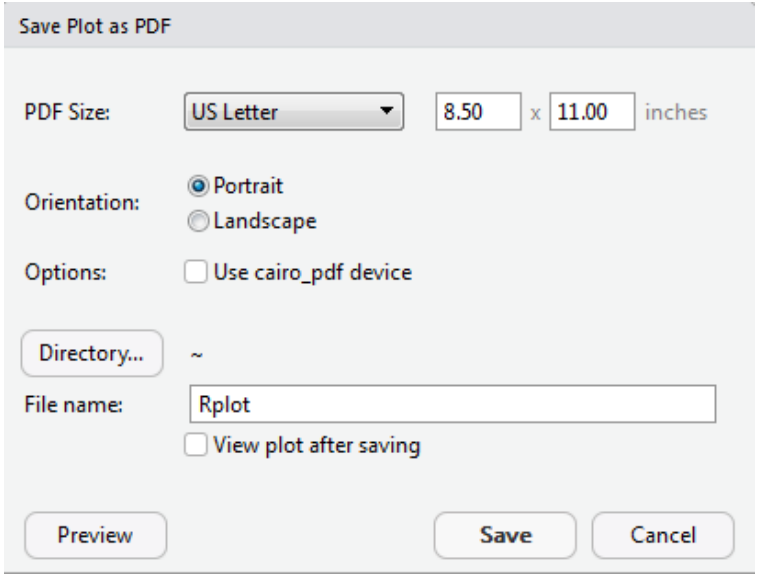


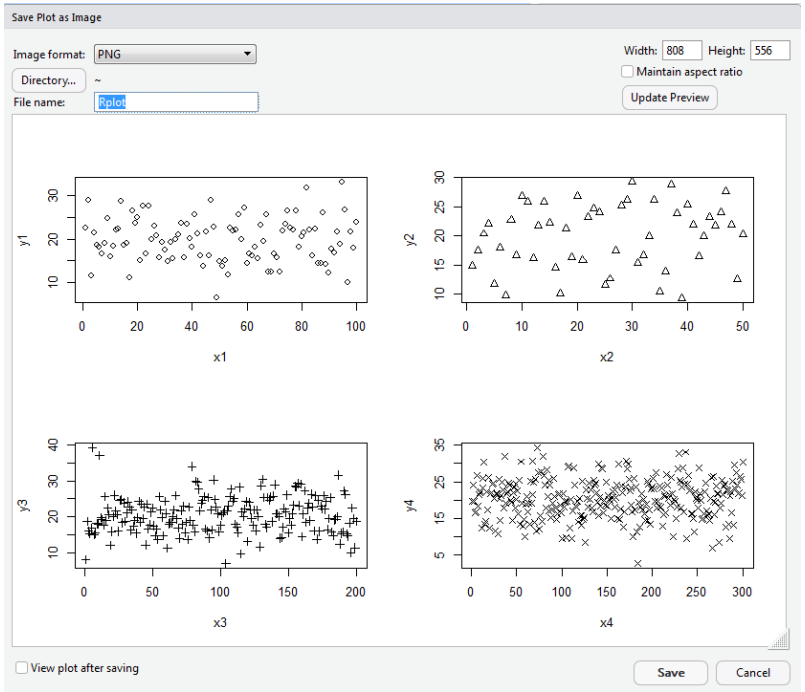




輸出圖形

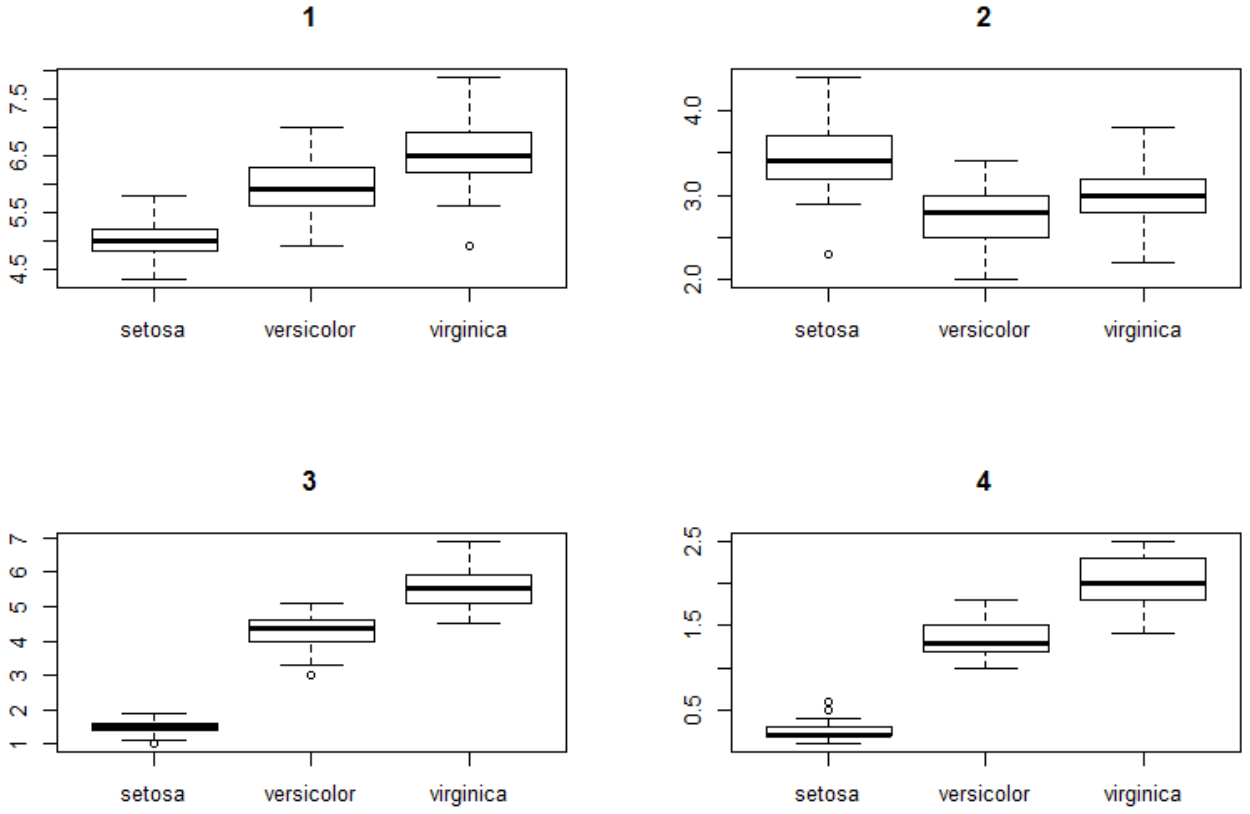

Zoom
Export
Publish
Save as Image...
Save as PDF...
Copy to
Save the current plot as a PDF file
y1
x1
y2
x2
y3
x3
y4
x4


Save Plot as PDF
PDF Size:
US Letter
8.50
x
11.00
inches
Orientation:
Portrait
Landscape
Options:
Use cairo_pdf device
Directory...
~
File name:
Rplot
View plot after saving
Preview
Save
Cancel


Save Plot as Image
Image format:
PNG
Directory...
~
File name:
Rplot
Width:
808
Height:
556
Maintain aspect ratio
Update Preview
View plot after saving
Save
Cancel

隨堂練習 1

1. 讀取 IRIS data
2. 畫出如以下圖形



Any Questions !?