R 資料分析應用：圖表繪製（二）

李智慎　副統計分析師

本次 eNews 將介紹的圖表有二維曲線圖、三維曲線圖、直方圖、長條圖、圓餅圖及盒鬚圖，同樣使用 CVD_ALL 這組資料作呈現，資料詳述內容定義可至 http://biostat.tmu.edu.tw/page/tmudata/help.docx 文件內觀看。

➢ 範例資料檔案

檔案可重網站網址 http://biostat.tmu.edu.tw/index.php/course/tmudata 下載，也可直接複製檔案路徑在`read.csv()`指令的參數`file`上，即可省去下載的步驟，直接連接雲端檔案使用，`stringsAsFactors`參數是防止`read.csv()`指另將字串欄位自動轉變為`factor`型態，在這裡們將檔案命名為`data.main`。

```
data.main <- read.csv(
  file = 'http://biostat.tmu.edu.tw/page/tmudata/CVD_All.csv',
  stringsAsFactors = FALSE
)
```

一開始一定要記得先觀看資料檢視，可以用`str()`指令觀看資料結構。

```
str(data.main)
```

可得到資料資訊包括資料總筆數 64489、總共 16 欄位和各欄位名稱型態及前幾筆資料。

```
'data.frame':   64489 obs. of  16 variables:
 $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 心血管疾病  : int  0 0 0 0 1 1 0 0 0 1 ...
 $ 年齡        : int  51 52 50 47 59 55 53 48 51 71 ...
 $ 性別        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ 追蹤時間    : int  1 1 3 5 5 3 4 4 3 5 ...
```

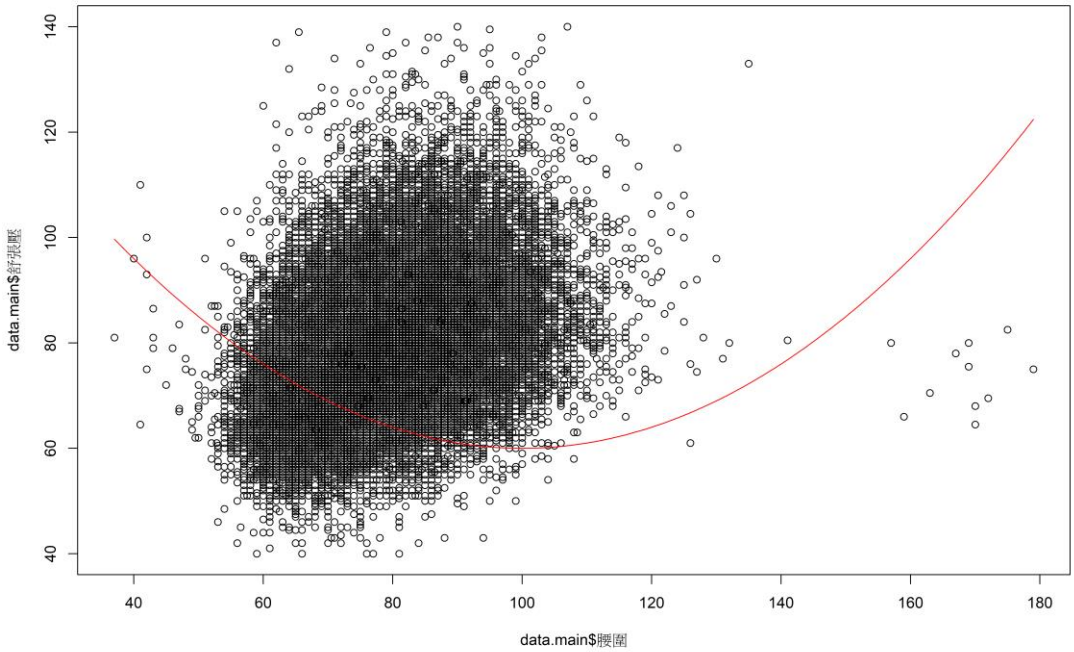
```
 $ 腰圍        : num  81 79 86.5 84 96 94 67 87 74 98 ...
 $ 收縮壓      : num  138 98 135 118 153 ...
 $ 舒張壓      : num  87 66 97 88.5 91.5 135 93 97.5 75 80 ...
 $ 空腹血糖    : int  194 101 90 88 90 200 148 98 75 NA ...
 $ 高密度脂蛋白: num  47 59 46 50 49 44 54.5 61.9 45 NA ...
 $ 三酸甘油酯  : int  517 186 153 201 132 995 220 112 169 NA ...
 $ 檳榔        : int  0 0 0 0 NA 0 0 0 0 1 ...
 $ 飲酒        : int  1 1 1 0 NA 1 1 0 1 1 ...
 $ 家族病史    : int  0 0 0 0 0 0 0 1 0 0 ...
 $ 抽菸        : int  1 1 1 0 1 0 0 0 1 1 ...
 $ 抽菸量      : int  2 2 1 0 1 0 0 0 1 2 ...
```

➢ 二維曲線圖

在這我們想用"腰圍"和"舒張壓"畫散佈圖，並在圖上畫一條線。首先創造了曲線函數`my.function()`，等同於$y = 0.01 \times (x - 100)^2 + 60$的曲線等式，在這利用了`min()`和`max()`找出腰圍資料的最小最大值，製造序列`x.seq`帶入`my.function()`得到`y.seq`，並用`lines()`指令即可繪製出曲線，程式碼如下

```
my.function <- function(x) {
  return(0.01*(x-100)^2 + 60)
}
x.seq <- seq(
  from = min(data.main$`腰圍`, na.rm = TRUE),
  to = max(data.main$`腰圍`, na.rm = TRUE),
  by = 0.5
)
y.seq <- my.function(x = x.seq)
plot(
  x = data.main$`腰圍`,
  y = data.main$`舒張壓`,
  type = "p"
)
lines(
```

```
  x = x.seq,
  y = y.seq,
  col = "red"
)
```

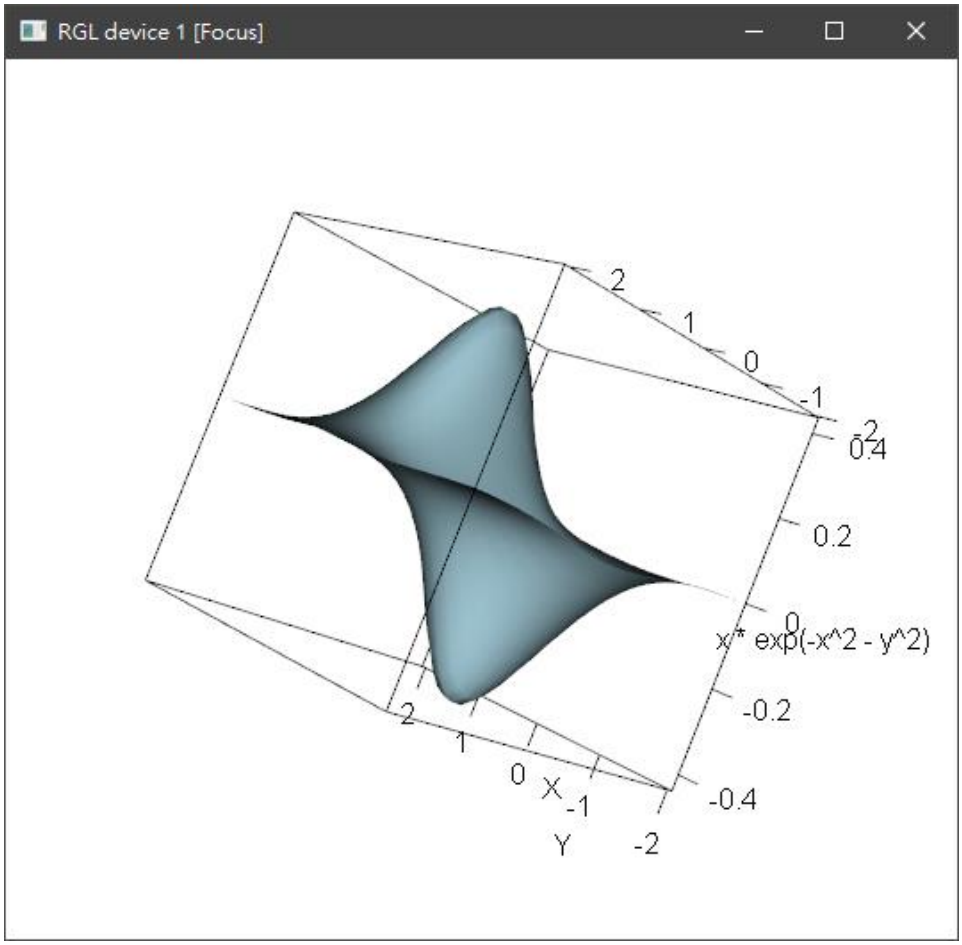


➢ 三維曲線圖

三圍曲線需使用套件`rgl`，製造三維空間圖首先需要有 XYZ 資料點才可畫出，在這我們製造`x.seq`和`y.seq`序列並且使用`outer()`帶入`my.function()`函數得到`z.seq`序列，接下來用`open3d()`開啟一個 R 畫布，最後用`persp3d()`帶入即可繪製完成，程式碼如下

```
library(rgl)
x.seq <- seq(
  from = -2,
  to = 2,
  length = 30
)
y.seq <- x.seq
my.function <- function(x, y) { x * exp(-x^2 - y^2) }
z.seq <- outer(
  X = x.seq,
```

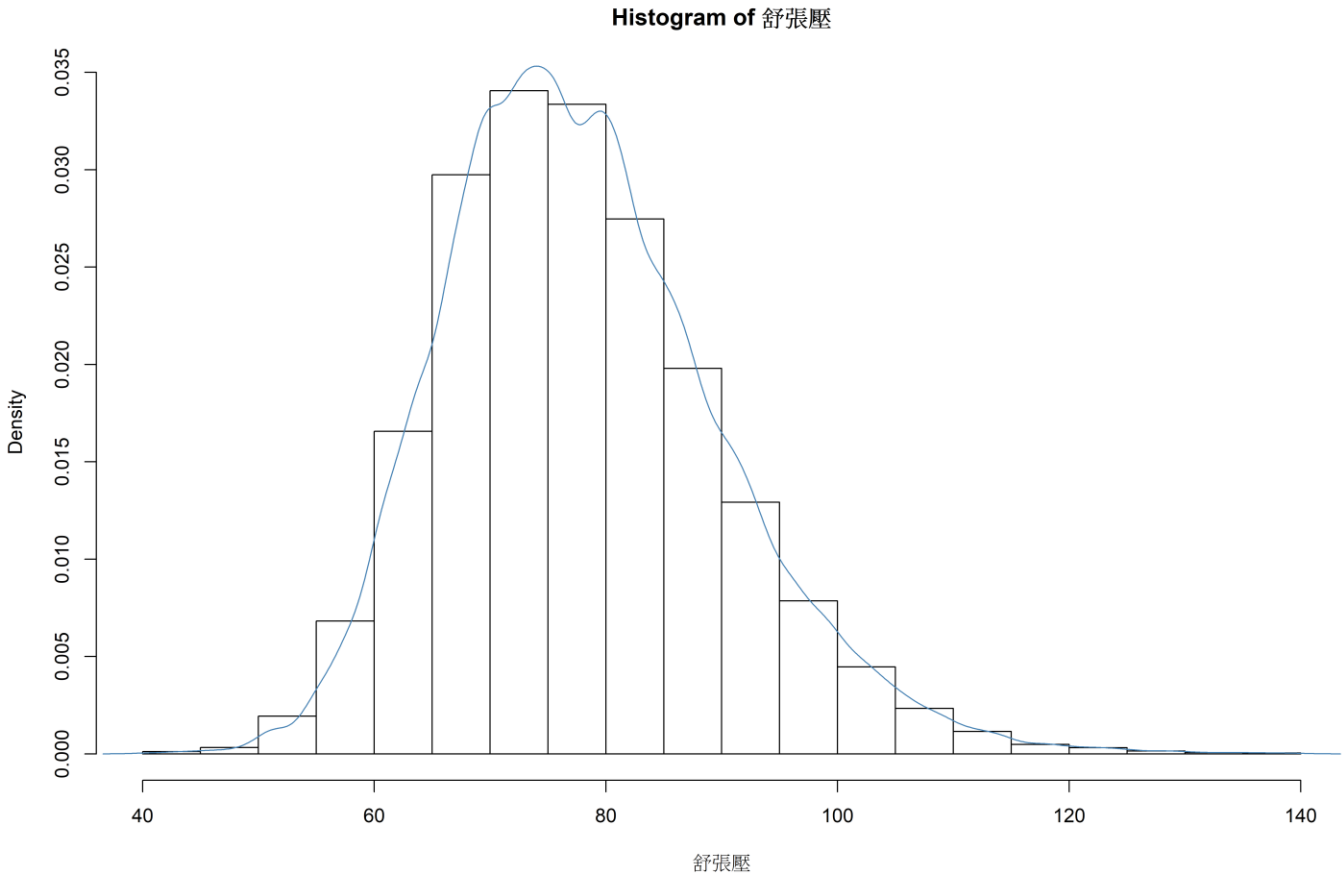
```
  Y = y.seq,
  FUN = my.function
)
open3d()
persp3d(
  x = x.seq,
  y = y.seq,
  z = z.seq,
  xlim = c(-2, 2),
  ylim = c(-2, 2),
  col = "lightblue",
  xlab = "X",
  ylab = "Y",
  zlab = "x * exp(-x^2 - y^2)"
)
```



➢ 直方圖

如果想知道資料的分布集中位置，用直方圖是一個很好的選擇，在 R 裡繪製非常簡單，只需將資料放置在`hist()`裡的`x`參數上即可，但在這裡我們多加了一條密度曲線在直方圖上，因此在這裡的`freq`參數需令為 FALSE，並且用`desity()`指令求出機率密度函數，接下來如前面例子一樣用`line()`指令即可加上機率密度曲線，程式碼如下

```
`舒張壓` <- na.exclude(data.main$`舒張壓`)
hist(
  x = `舒張壓`,
  freq = FALSE
)
my.density <- density(`舒張壓`)
lines(my.density, col = "steelblue")
```



➢ 長條圖

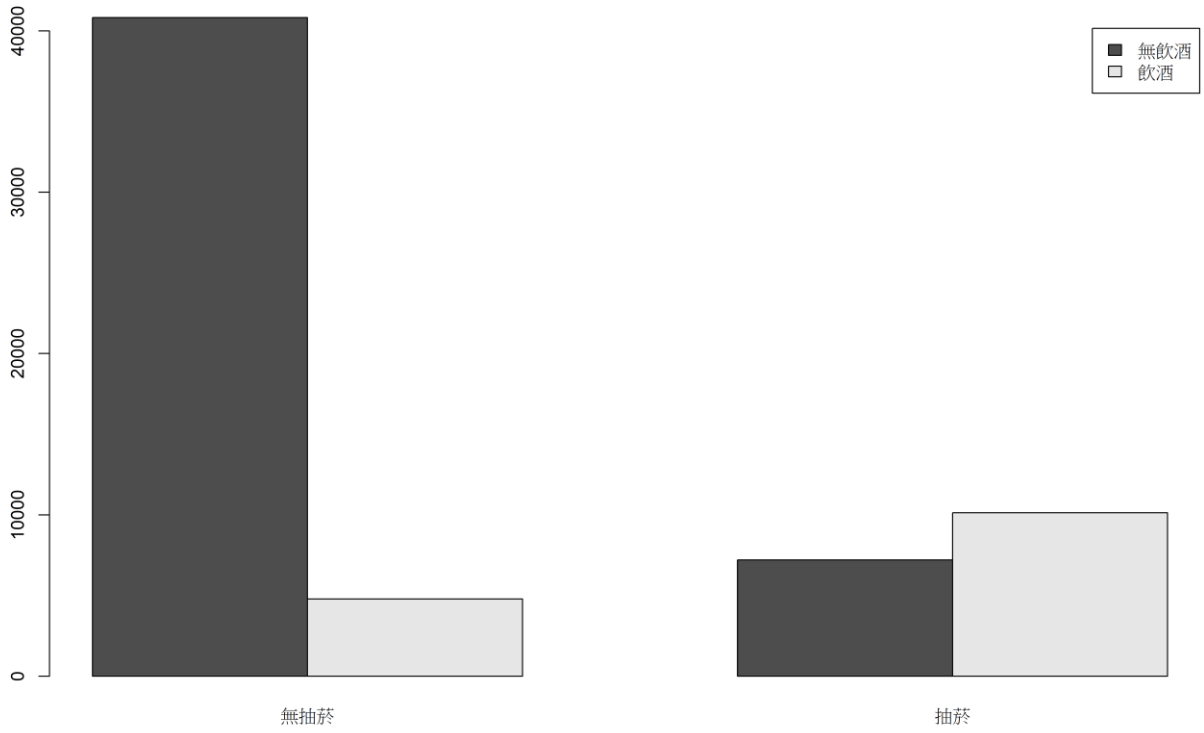
長條圖是很好得知資料個數的呈現方式，在這我們觀看資料內抽菸和飲酒的人數，並用長條圖呈現出來，在畫圖之前必須先計算出個數，在 R 裡只需用`table()`即可計算出，再用`rownames()`和`colnames()`函式修改一下名稱即可得到如下表格

```
my.table <- table(data.main[,c('飲酒', '抽菸')])
rownames(my.table) <- c('無飲酒', '飲酒')
colnames(my.table) <- c('無抽菸', '抽菸')
```

	無抽菸	抽菸
無飲酒	40831	7205
飲酒	4781	10127

再將資料帶入`barplot()`內即可得到長條圖如下

```
barplot(
  height = my.table,
  beside = TRUE,
  legend = rownames(my.table)
)
```

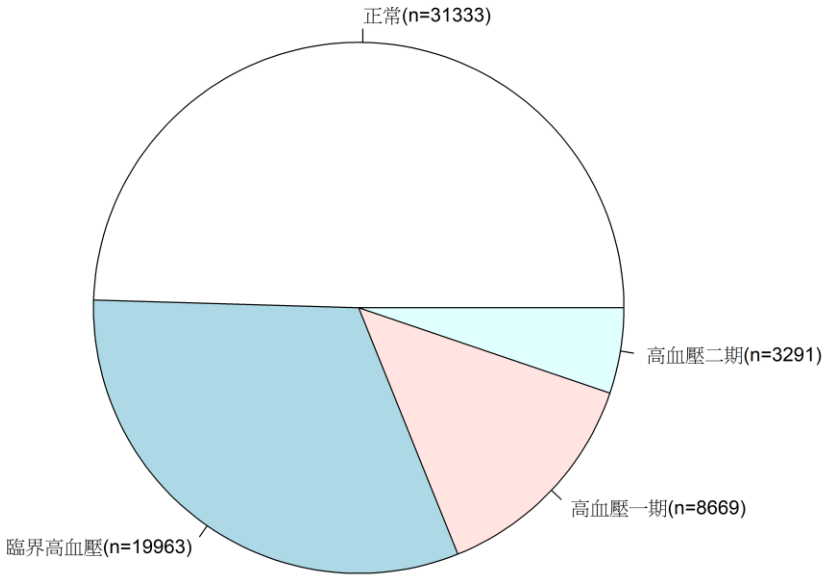


➢ 圓餅圖

圓餅圖是很常見的一個圖形用來得知各類別個數，但有一缺點是當種類較多時，會較難從各扇型的弧度得知各類別的多寡，在人類視覺上要得知資料多寡，長條圖還是最為清楚且較不受種類總數限制的圖形，在這用的是"收縮壓"欄位繪製圓餅圖，我們依照收縮壓的高低用`cut()`先將資料切割成正常、臨界高血壓、高血壓一期和高血壓二期 4 個類別，切割點是參考維基百科 https://zh.wikipedia.org/wiki/血壓 上的資訊，切割完後一樣使用`table()`計算個數，帶入`pie()`內即可繪製圓餅圖，稍微特殊的是我用利用`paste0()`製造了類別個數字串序列，並附帶在`labels`參數上讓圓餅圖能呈現出來，程式碼如下

```
`收縮壓` <- cut(
  x = na.exclude(data.main$`收縮壓`),
  breaks = c(-Inf, 120, 140, 160, Inf),
  labels = c("正常", "臨界高血壓", "高血壓一期", "高血壓二期")
)
my.table <- table(`收縮壓`)
pie(
  x = my.table,
```

```
  labels = paste0(
    names(my.table),
    '(n=',
    as.numeric(my.table),
    ')'
  )
)
```

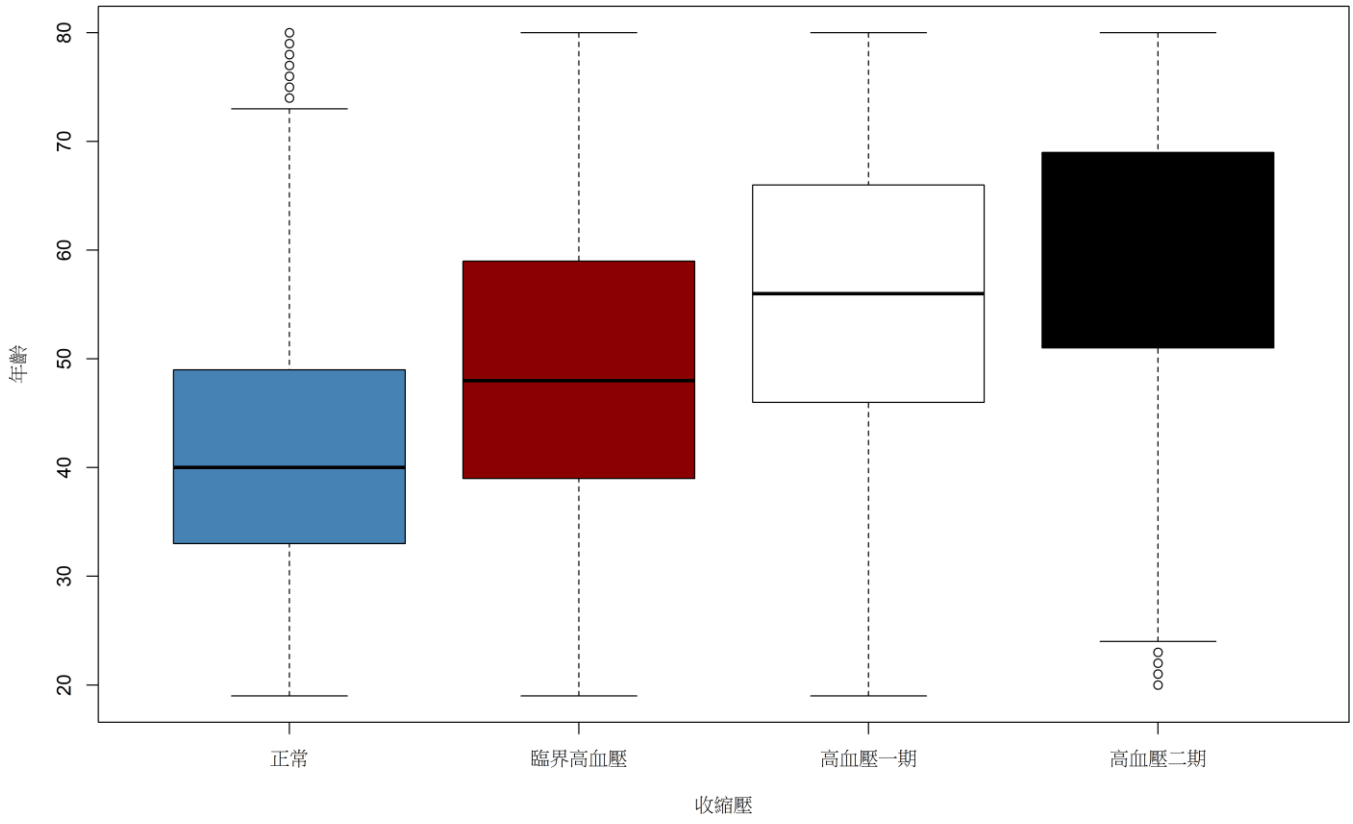


- 盒鬚圖

在敘述統計裡還有一項常需要得知的就是分位數，而盒鬚圖就是一個很好觀看分位數的圖形，在這我們利用前例分類好的收縮壓，觀看在各類收縮壓情況下的年齡高低狀況，在 R 用`boxplot()`即可繪製出盒鬚圖，比較需注意的是`formula`參數內，在'~'符號左邊放置的是數值右邊放置的是類別，在此例分別是年齡和收縮壓，程式碼如下

```
my.data <- na.exclude(data.main)
my.data$`收縮壓` <- cut(
```

```
  x = my.data$`收縮壓`,
  breaks = c(-Inf, 120, 140, 160, Inf),
  labels = c("正常", "臨界高血壓", "高血壓一期", "高血壓二期")
)
boxplot(
  formula = `年齡` ~ `收縮壓`,
  data = my.data,
  col = c(
    'steelblue',
    'darkred',
    'white',
    'black'
  ),
  xlab = "收縮壓",
  ylab = "年齡"
)
```



➢ 參考資料

1. R 軟體 應用統計方法 陳景祥編著 東華書局
2. Package 'rgl' - CRAN-R - R Projecthttps://cran.r-project.org/web/packages/rgl/rgl.pdf
3. 維基百科血壓 https://zh.wikipedia.org/wiki/%E8%A1%80%E5%A3%93