

106-2: EE4052

通識課程： 計算機程式設計 之旅

Computer Programming

# Unit 13: 探索性資料分析

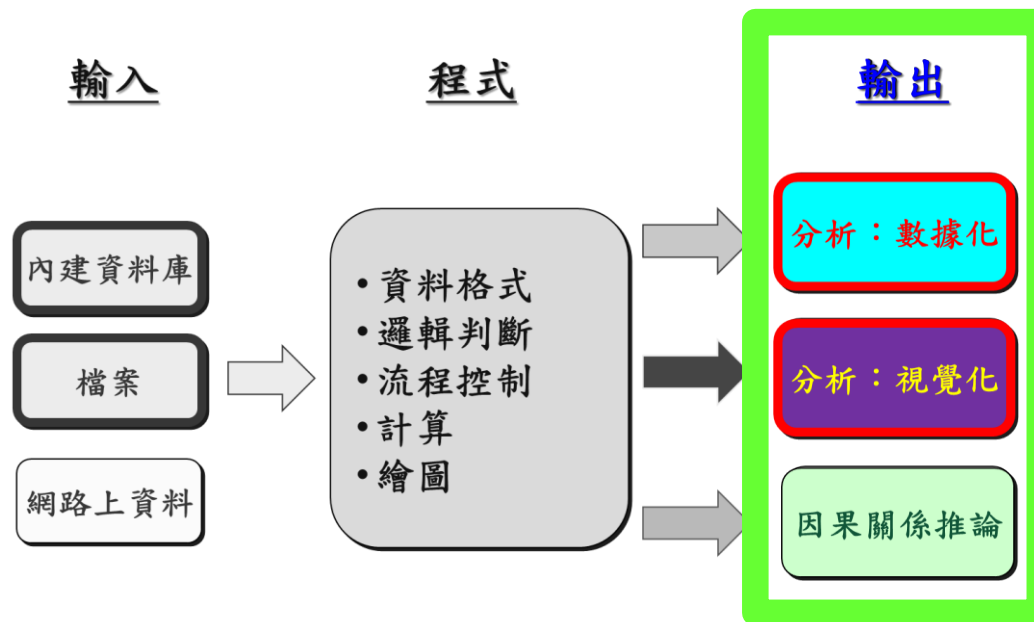
連 豐 力

臺大電機系

Feb 2018 - Jun 2018

# 課程主題進度

- **U01:** 課程介紹：討論主題，作業，報告，進行方式
- **U02:** 主題，案例，程式，演算法，資源
- **U03:** 設定軟體 **R** 與 **Rstudio**
- **U04:** 數據處理與繪圖指令功能
- **U05:** 資料類別與基本運算
- **U06:** 邏輯判斷與流程控制
- **U07:** 函數：計算與排序
- **U08:** 多維度資料格式
- **U09:** 檔案資料輸入與輸出
- **U10:** 繪圖功能與文字
- **U11:** 多重繪圖與顏色
- **U12:** 資料間的相關性
- **U13:** 探索性資料分析
- **U14:** 資料連結分析
- **U15:** 影像與動畫



# 大綱

## ■ 數據化探索

- 變數概況
- 變數詳情
- 分布指標

## ■ 視覺化探索

- 長條圖
- 累積分布圖
- 箱形圖 盒鬚圖
- 橫條圖
- 點陣圖
- 圓形圖

# 資料庫

- **Insurance 資料集：**  
記錄了某保險公司在 1973 年第三季 **車險投保人的** 相關資料
- **District:** 投保人家庭住址所在區域，  
設定值：**1-4**
- **Group:** 所投保汽車的引擎排氣量：  
**1升，1-1.5升，1.5-2升，大於2升**，四個等級
- **Age:** 投保人年齡：  
**小於25歲，25-29歲，30-35歲，大於35歲**，四組別
- **HOLDERS:** 投保人**數量**
- **Claims:** 要求索賠的投保人**數量**

- `install.packages( "MASS" )`
- `library( MASS )`
- `data( Insurance )`
- `nrow( Insurance )`
- `ncol( Insurance )`
- `dim( Insurance )`
- `head( Insurance )`
- `tail( Insurance )`

```
> head( Insurance )
```

	District	Group	Age	Holders	Claims
1	1	<1	<25	197	38
2	1	<1	25-29	264	35
3	1	<1	30-35	246	20
4	1	<1	>35	1680	156
5	1	1-1.5	<25	284	63
6	1	1-1.5	25-29	536	84

```
> tail( Insurance )
```

	District	Group	Age	Holders	Claims
59	4	1.5-2	30-35	68	16
60	4	1.5-2	>35	344	63
61	4	>2	<25	3	0
62	4	>2	25-29	16	6
63	4	>2	30-35	25	8
64	4	>2	>35	114	33

# 數據化探索 - 變數概況

# 變數概況

- # variable attribute, 資料集變數屬性

- attributes( Insurance )

```
> attributes( Insurance )
```

```
$names
```

```
[1] "District" "Group" "Age" "Holders" "Claims"
```

```
$class
```

```
[1] "data.frame"
```

```
$row.names
```

```
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
[20] 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
[39] 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
[58] 58 59 60 61 62 63 64
```

- # internal structure, 內部結構

- str( Insurance )

```
> str( Insurance )
```

```
'data.frame':      64 obs. of  5 variables:
 $ District: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ Group   : Ord.factor w/ 4 levels "<11"<"1-1.51"<..: 1 1 1 1 2 2 2 2 3 3 ...
 $ Age     : Ord.factor w/ 4 levels "<25"<"25-29"<..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Holders : int   197 264 246 1680 284 536 696 3582 133 286 ...
 $ Claims  : int    38 35 20 156 63 84 89 400 19 52 ...
```



# 變數概況

- # summary, 統計指標值
- summary( Insurance )

> summary( Insurance )

District	Group	Age	Holders	Claims
1:16	<17 :16	<25 :16	Min. : 3.00	Min. : 0.00
2:16	1-1.57:16	25-29:16	1st Qu.: 46.75	1st Qu.: 9.50
3:16	1.5-21:16	30-35:16	Median : 136.00	Median : 22.00
4:16	>21 :16	>35 :16	Mean : 364.98	Mean : 49.23
			3rd Qu.: 327.50	3rd Qu.: 55.50
			Max. : 3582.00	Max. : 400.00

- 定性變數：各水準的設定值的頻數
  - District, Group, Age: 四種的出現頻率都是16個
- 定量變數：統計數字指標
  - 最小值，第一四分位點，中位數（第二四分位點），平均值，第三四分位點，最大值
  - 中位數與平均值的差異：判斷資料的偏倚程度，左偏或右偏

# 數據化探索 - 變數詳情

# describe( ) in Hmisc, 更詳細的變數情況，使用 Hmisc 軟體套件 describe( )

```
install.packages( "Hmisc" )
```

```
library( Hmisc )
```

```
describe( Insurance[ , 1:3] )
```

n: 樣本總個數

missing: 缺失樣本數

unique: 水準個數

每一個水準：

設定值，頻數，頻率

```
> describe( Insurance[ , 1:3] )
```

```
Insurance[, 1:3]
```

```
3 Variables      64 Observations
```

```
-----  
District
```

```
  n missing distinct  
64      0         4
```

```
1 (16, 0.25), 2 (16, 0.25), 3 (16, 0.25), 4 (16, 0.25)
```

```
-----  
Group
```

```
  n missing distinct  
64      0         4
```

```
<11 (16, 0.25), 1-1.51 (16, 0.25), 1.5-21 (16, 0.25), >21 (16, 0.25)
```

```
-----  
Age
```

```
  n missing distinct  
64      0         4
```

```
<25 (16, 0.25), 25-29 (16, 0.25), 30-35 (16, 0.25), >35 (16, 0.25)
```

# describe( ) in Hmisc, 更詳細的變數情況，使用 Hmisc 軟體套件 describe( )

```
install.packages( "Hmisc" )
```

```
library( Hmisc )
```

```
describe( Insurance[ , 4:5] )
```

```
> describe( Insurance[ , 4:5] )
```

```
Insurance[, 4:5]
```

```
2 Variables      64 Observations
```

```
-----  
Holders
```

n	missing	distinct	Info	Mean	Gmd	.05
64	0	63	1	365	497.1	16.30
.10	.25	.50	.75	.90	.95	
24.00	46.75	136.00	327.50	868.90	1639.25	

```
lowest : 3 7 9 16 18, highest: 1635 1640 1680 2443 3582
```

```
-----  
Claims
```

n	missing	distinct	Info	Mean	Gmd	.05
64	0	46	0.999	49.23	60.66	3.15
.10	.25	.50	.75	.90	.95	
4.30	9.50	22.00	55.50	101.70	182.35	

```
lowest : 0 2 3 4 5, highest: 156 187 233 290 400
```

n: 樣本總個數

missing: 缺失樣本數

unique: 水準個數

每一個水準：

設定值，頻數，頻率

# `basicStats()` in `fBasics`, 輸出指標更豐富，使用 `fBasics` 軟體套件

```
install.packages("fBasics")
```

```
library(fBasics)
```

```
basicStats(Insurance$Holders)
```

```
> basicStats(Insurance$Holders)
```

觀測樣本數

遺漏數

最小

最大

第一四分位點

第三四分位點

平均值

中位數

總和

標準誤差平均值

95% 置信水準平均值之置信下限

95% 置信水準平均值之置信上限

變異量

標準誤差

偏度

峰度

X..Insurance.Holders

nobs	6.400000e+01
NA's	0.000000e+00
Minimum	3.000000e+00
Maximum	3.582000e+03
1. Quartile	4.675000e+01
3. Quartile	3.275000e+02
Mean	3.649844e+02
Median	1.360000e+02
Sum	2.335900e+04
SE Mean	7.784632e+01
LCL Mean	2.094209e+02
UCL Mean	5.205478e+02
Variance	3.878432e+05
Stdev	6.227706e+02
Skewness	3.127833e+00
Kurtosis	1.099961e+01

# 數據化探索 - 分布指標

# 分布指標 – Insurance

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
install.packages( "timeDate" )
```

```
library( timeDate )
```

```
skewness( Insurance[ , 4:5] )
```

```
kurtosis( Insurance[ , 4:5] )
```

```
> skewness( Insurance[ , 4:5] )
```

```
  Holders  Claims  
3.127833 2.877292
```

```
> kurtosis( Insurance[ , 4:5] )
```

```
  Holders  Claims  
10.999610 9.377258
```

# 分布指標 - Insurance

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

**偏度**： 衡量資料的偏倚程度或對稱程度

= 0: 正態分布，完全對稱

[-1, 1]: 對稱性較強，不存在左偏或右偏

> 1: 右偏

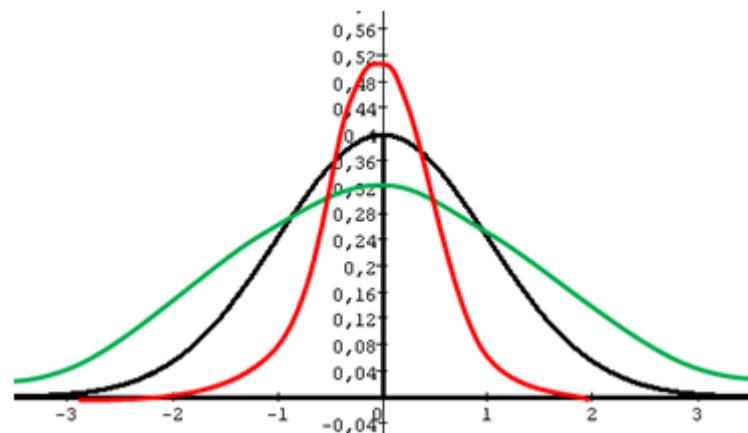
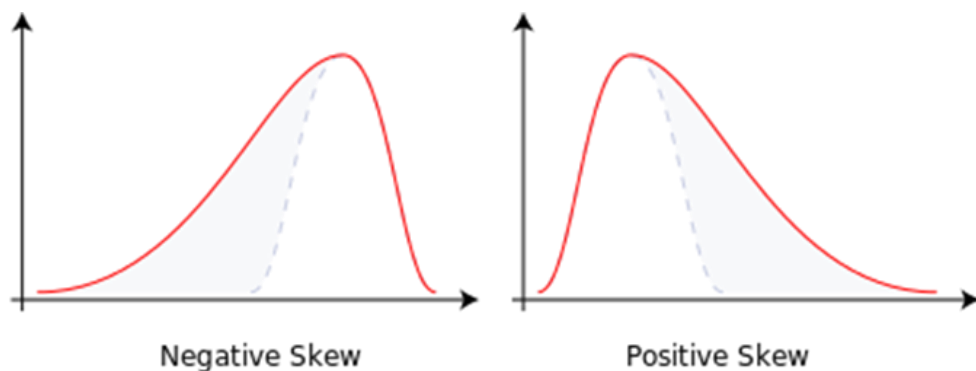
< -1: 左偏

**峰度**： 衡量資料的分布型態的陡緩程度，集中或分散

= 0: 集散程度與正態分布相同，為標準峰度

> 0: 比正態分布較為陡峭，為尖頂峰度

< 0: 比正態分布較為平坦，為平頂峰度





# 分布指標 – Insurance

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
set.seed(1)
```

```
s1 <- .Random.seed
```

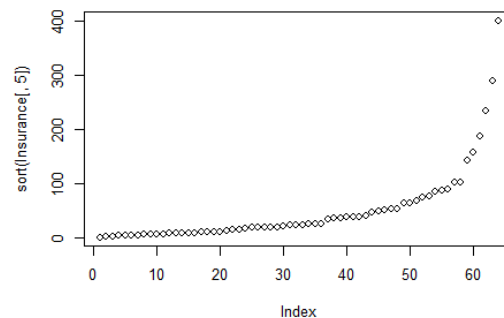
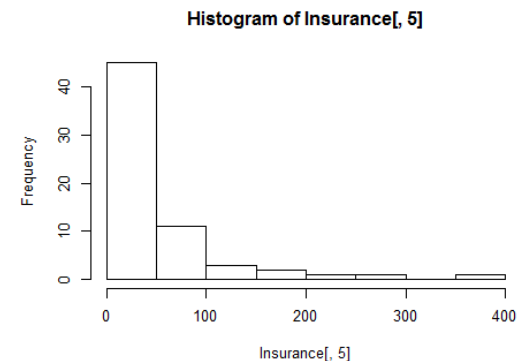
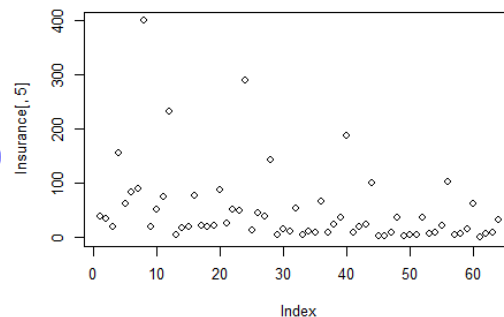
```
y <- rnorm( 100 )
```

```
layout( matrix( 1:4, nrow=2 ) )
```

```
plot( y )
```

```
plot( sort( y ) )
```

```
hist( y )
```



# 分布指標 – Insurance

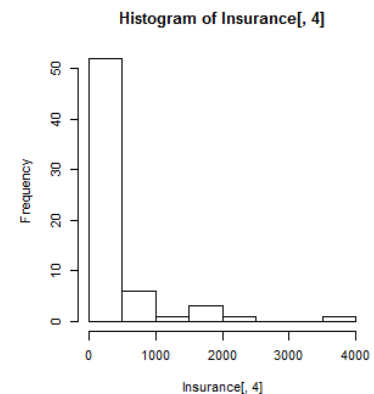
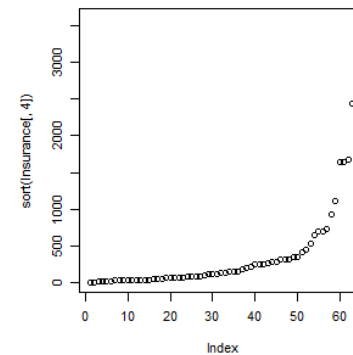
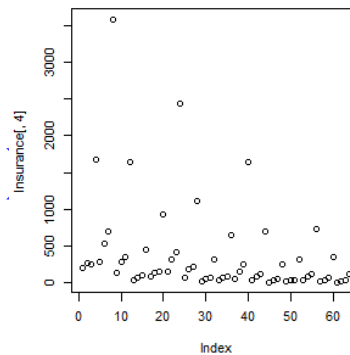
# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
layout( matrix( 1:6, nrow=2, byrow=T ) )
```

```
plot( Insurance[ , 4 ] )
```

```
plot( sort( Insurance[ , 4 ] ) )
```

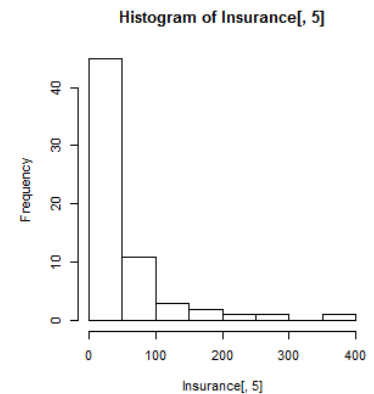
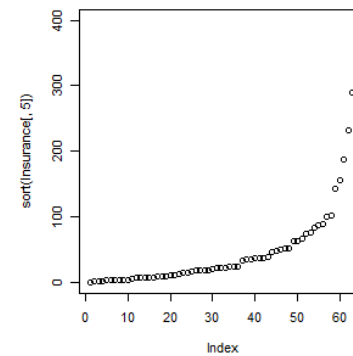
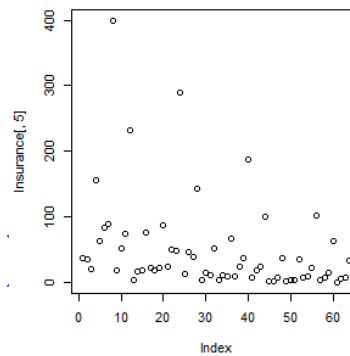
```
hist( Insurance[ , 4 ] )
```



```
plot( Insurance[ , 5 ] )
```

```
plot( sort( Insurance[ , 5 ] ) )
```

```
hist( Insurance[ , 5 ] )
```



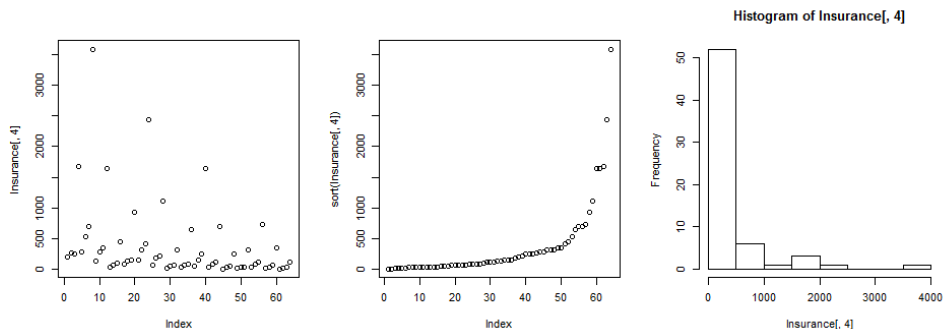
# 分布指標 - Insurance

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

- 偏度 :** 衡量資料的偏倚程度或對稱程度
- = 0: 正態分布，完全對稱
  - [-1, 1]: 對稱性較強，不存在左偏或右偏
  - > 1: 右偏
  - < -1: 左偏
- 峰度 :** 衡量資料的分布型態的陡緩程度，集中或分散
- = 0: 集散程度與正態分布相同，為標準峰度
  - > 0: 比正態分布較為陡峭，為尖頂峰度
  - < 0: 比正態分布較為平坦，為平頂峰度

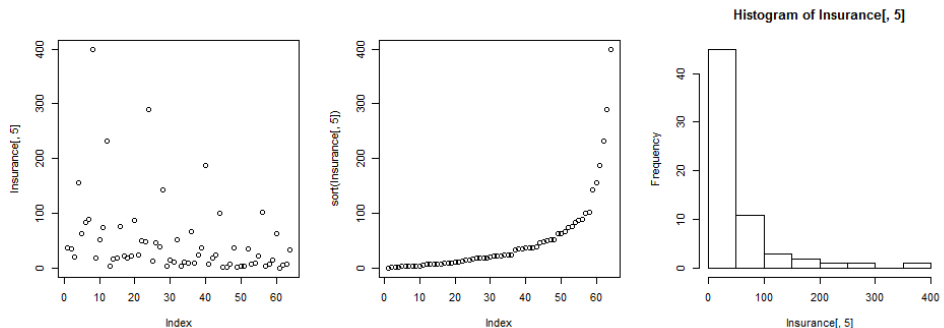
> skewness( Insurance[ , 4:5] )

holders	claims
3.127833	2.877292



> kurtosis( Insurance[ , 4:5] )

holders	claims
10.999610	9.377258



# 分布指標 - iris

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
layout( matrix( 1:8, nrow=2, byrow=F ) )
```

```
plot( sort( iris[ , 1 ] ) )
```

```
hist( iris[ , 1 ] )
```

```
plot( sort( iris[ , 2 ] ) )
```

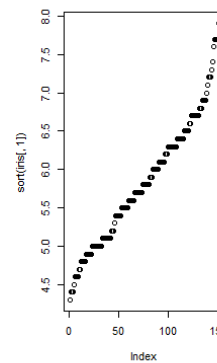
```
hist( iris[ , 2 ] )
```

```
plot( sort( iris[ , 3 ] ) )
```

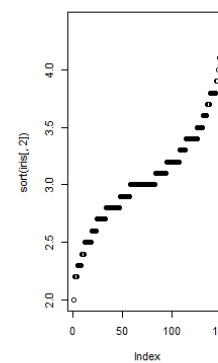
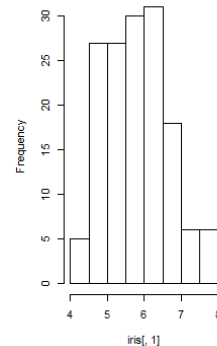
```
hist( iris[ , 3 ] )
```

```
plot( sort( iris[ , 4 ] ) )
```

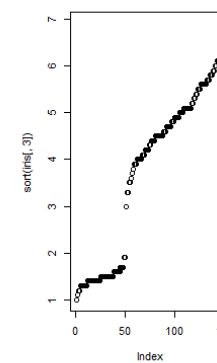
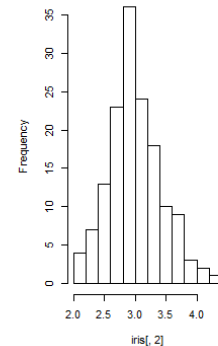
```
hist( iris[ , 4 ] )
```



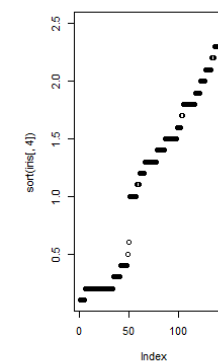
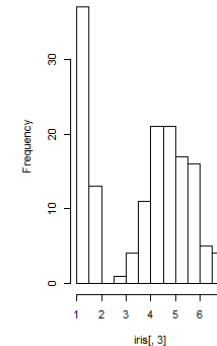
Histogram of iris[, 1]



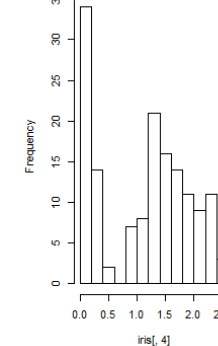
Histogram of iris[, 2]



Histogram of iris[, 3]



Histogram of iris[, 4]



# 分布指標 - iris

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

install.packages( "timeDate" )

library( timeDate )

skewness( iris[ , 1:4] )

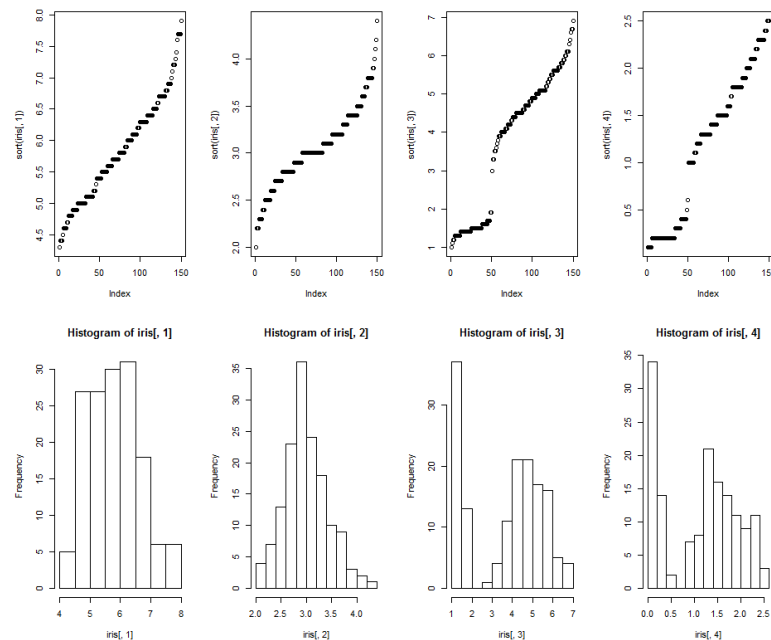
kurtosis( iris[ , 1:4] )

> skewness( iris[ , 1:4] )

Sepal.Length	Sepal.width	Petal.Length	Petal.Width
0.3086407	0.3126147	-0.2694109	-0.1009166

> kurtosis( iris[ , 1:4] )

Sepal.Length	Sepal.width	Petal.Length	Petal.Width
-0.6058125	0.1387047	-1.4168574	-1.3581792



**偏度**：衡量資料的偏倚程度或對稱程度

- = 0: 正態分布，完全對稱
- [-1, 1]: 對稱性較強，不存在左偏或右偏
- > 1: 右偏
- < -1: 左偏

**峰度**：衡量資料的分布型態的陡緩程度，集中或分散

- = 0: 集散程度與正態分布相同，為標準峰度
- > 0: 比正態分布較為陡峭，為尖頂峰度
- < 0: 比正態分布較為平坦，為平頂峰度

# 分布指標 – Typhoon-01

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
tphdata <- read.table( "L:/DataWD/Typhoon-01.txt", header = TRUE )
```

```
layout( matrix( 1:8, nrow=2, byrow=F ) )
```

```
plot( sort( tphdata[ , 1 ] ) )
```

```
hist( tphdata[ , 1 ] )
```

```
plot( sort( tphdata[ , 2 ] ) )
```

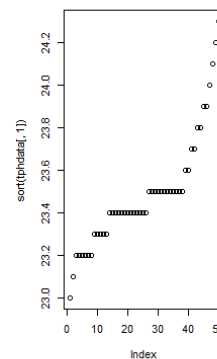
```
hist( tphdata[ , 2 ] )
```

```
plot( sort( tphdata[ , 3 ] ) )
```

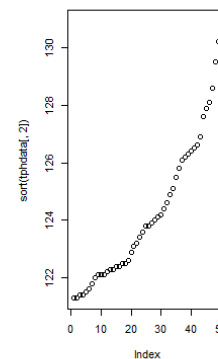
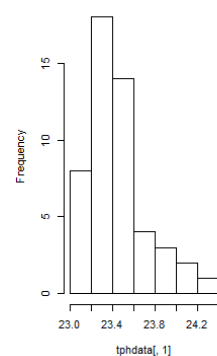
```
hist( tphdata[ , 3 ] )
```

```
plot( sort( tphdata[ , 4 ] ) )
```

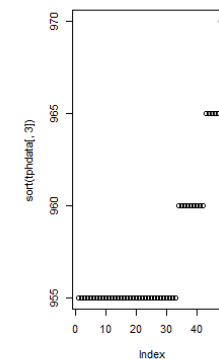
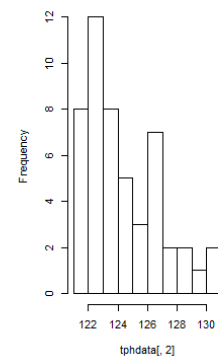
```
hist( tphdata[ , 4 ] )
```



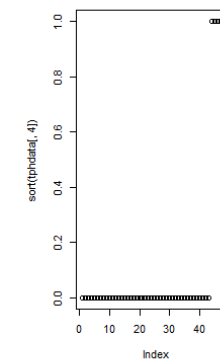
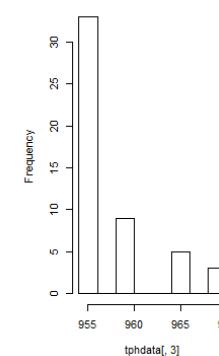
Histogram of tphdata[, 1]



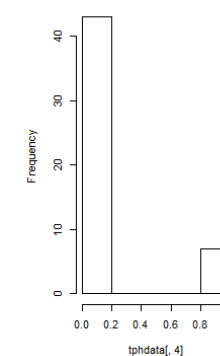
Histogram of tphdata[, 2]



Histogram of tphdata[, 3]



Histogram of tphdata[, 4]



# 分布指標 – Typhoon-01

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
install.packages( "timeDate" )
```

```
library( timeDate )
```

```
skewness( tphdata[ , 1:4] )
```

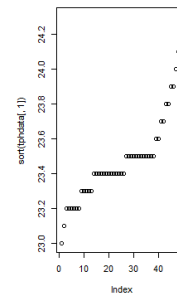
```
kurtosis( tphdata[ , 1:4] )
```

```
> skewness( tphdata[ , 1:4] )
```

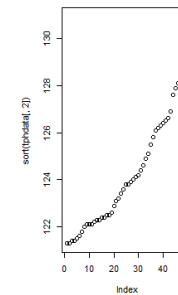
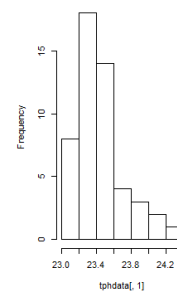
x1	x2	x3	x4
1.0759446	0.8048285	1.4330051	2.0130676

```
> kurtosis( tphdata[ , 1:4] )
```

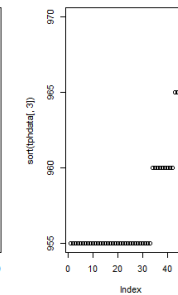
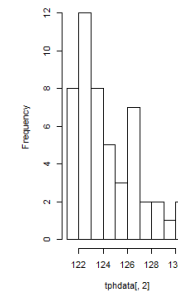
x1	x2	x3	x4
1.0076500	-0.2838985	0.8818208	2.0955442



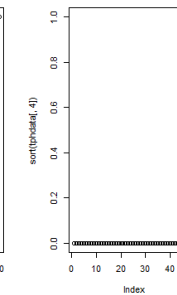
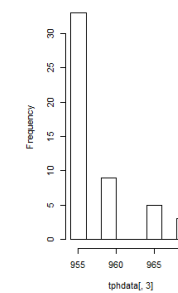
Histogram of tphdata[, 1]



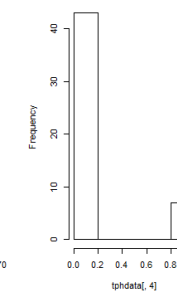
Histogram of tphdata[, 2]



Histogram of tphdata[, 3]



Histogram of tphdata[, 4]



**偏度**：衡量資料的偏倚程度或對稱程度

- = 0: 正態分布，完全對稱
- [-1, 1]: 對稱性較強，不存在左偏或右偏
- > 1: 右偏
- < -1: 左偏

**峰度**：衡量資料的分布型態的陡緩程度，集中或分散

- = 0: 集散程度與正態分布相同，為標準峰度
- > 0: 比正態分布較為陡峭，為尖頂峰度
- < 0: 比正態分布較為平坦，為平頂峰度

# 分布指標 – Typhoon-01

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
tphdata <- read.table( "L:/DataWD/Typhoon-01.txt", header = TRUE )
```

```
layout( matrix( 1:8, nrow=2, byrow=F ) )
```

```
plot( sort( tphdata[ , 5 ] ) )
```

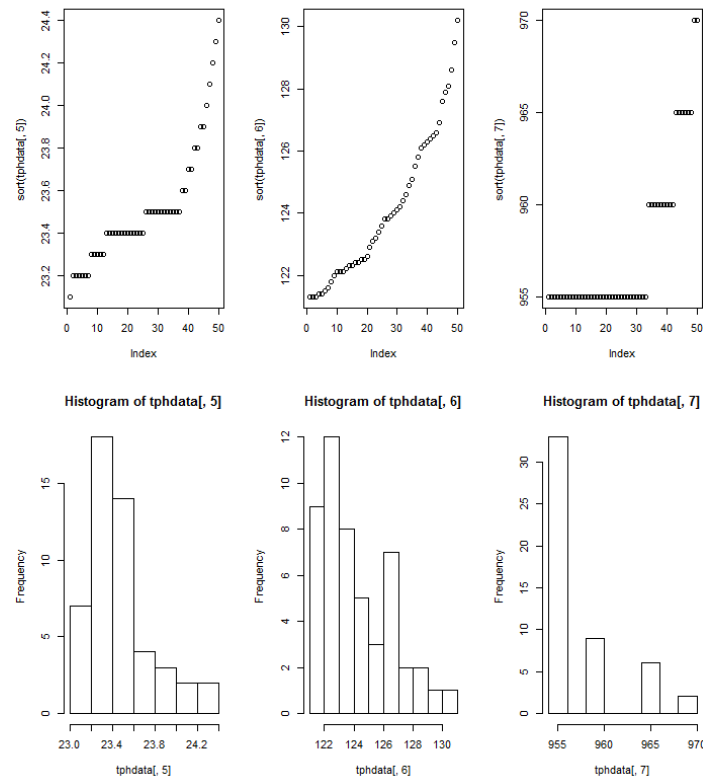
```
hist( tphdata[ , 5 ] )
```

```
plot( sort( tphdata[ , 6 ] ) )
```

```
hist( tphdata[ , 6 ] )
```

```
plot( sort( tphdata[ , 7 ] ) )
```

```
hist( tphdata[ , 7 ] )
```





# 分布指標 – Typhoon-01

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

install.packages( "timeDate" )

library( timeDate )

skewness( tphdata[ , 5:7] )

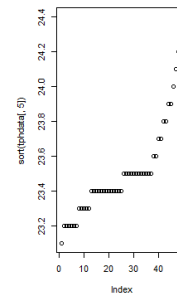
kurtosis( tphdata[ , 5:7] )

> skewness( tphdata[ , 5:7] )

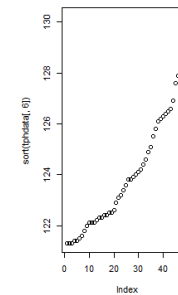
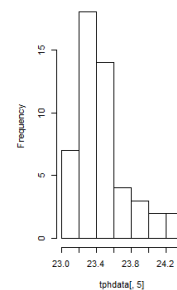
d1	d2	d3
1.2903782	0.7363766	1.3772497

> kurtosis( tphdata[ , 5:7] )

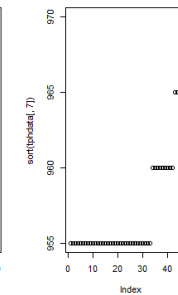
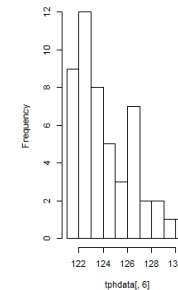
d1	d2	d3
1.2025491	-0.4650986	0.7547749



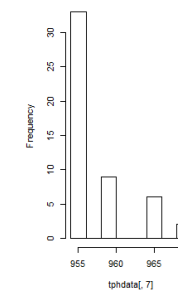
Histogram of tphdata[, 5]



Histogram of tphdata[, 6]



Histogram of tphdata[, 7]



**偏度**：衡量資料的偏倚程度或對稱程度

= 0: 正態分布，完全對稱

[-1, 1]: 對稱性較強，不存在左偏或右偏

> 1: 右偏

< -1: 左偏

**峰度**：衡量資料的分布型態的陡緩程度，集中或分散

= 0: 集散程度與正態分布相同，為標準峰度

> 0: 比正態分布較為陡峭，為尖頂峰度

< 0: 比正態分布較為平坦，為平頂峰度

# 分布指標 - CO2

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

CO2

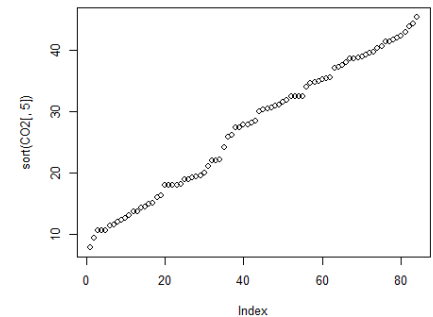
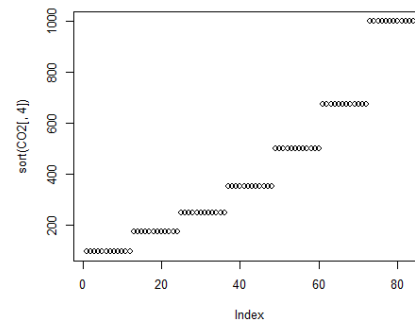
```
layout( matrix( 1:4, nrow=2, byrow=F ) )
```

```
plot( sort( CO2[ , 4 ] ) )
```

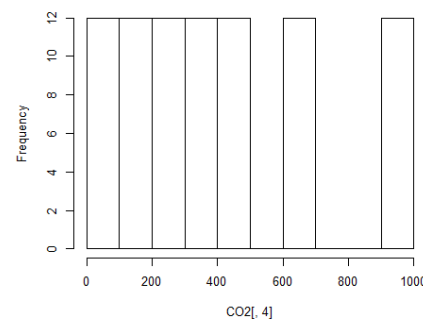
```
hist( CO2[ , 4 ] )
```

```
plot( sort( CO2[ , 5 ] ) )
```

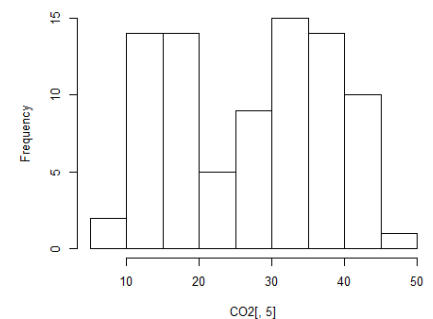
```
hist( CO2[ , 5 ] )
```



Histogram of CO2[, 4]



Histogram of CO2[, 5]



# 分布指標 - CO2

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
install.packages("timeDate")
```

```
library(timeDate)
```

```
skewness(CO2[, 4:5])
```

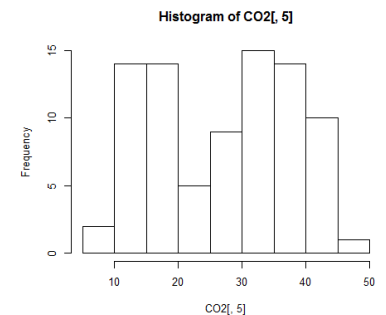
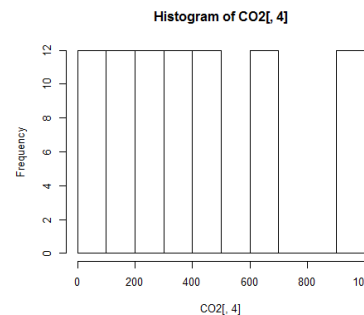
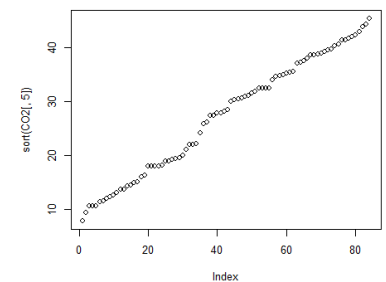
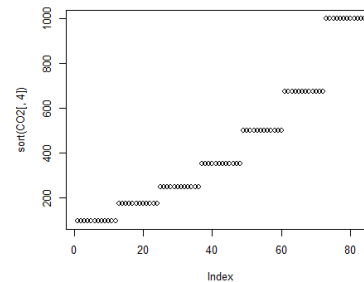
```
kurtosis(CO2[, 4:5])
```

```
> skewness(CO2[, 4:5])
```

```
      conc      uptake  
0.7201458 -0.1040551
```

```
> kurtosis(CO2[, 4:5])
```

```
      conc      uptake  
-0.6826587 -1.3482674
```



**偏度**：衡量資料的偏倚程度或對稱程度

= 0: 正態分布，完全對稱

[-1, 1]: 對稱性較強，不存在左偏或右偏

> 1: 右偏

< -1: 左偏

**峰度**：衡量資料的分布型態的陡緩程度，集中或分散

= 0: 集散程度與正態分布相同，為標準峰度

> 0: 比正態分布較為陡峭，為尖頂峰度

< 0: 比正態分布較為平坦，為平頂峰度

# 分布指標 – weather

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

```
install.packages( "rattle.data" )
```

```
library( rattle.data )
```

```
layout( matrix( 1:8, nrow=2, byrow=F ) )
```

```
plot( sort( weather[ , 3 ] ) )
```

```
hist( weather[ , 3 ] )
```

```
plot( sort( weather[ , 4 ] ) )
```

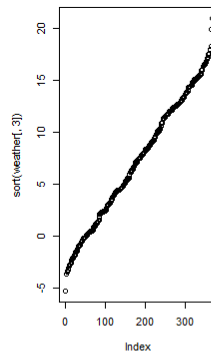
```
hist( weather[ , 4 ] )
```

```
plot( sort( weather[ , 5 ] ) )
```

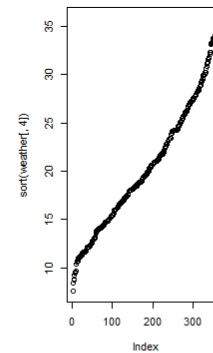
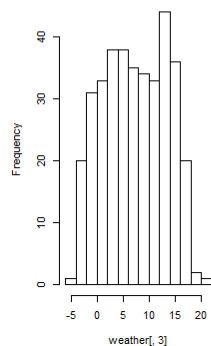
```
hist( weather[ , 5 ] )
```

```
plot( sort( weather[ , 6 ] ) )
```

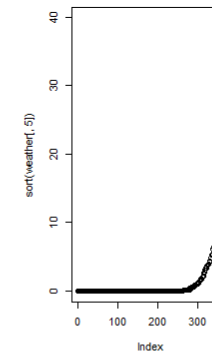
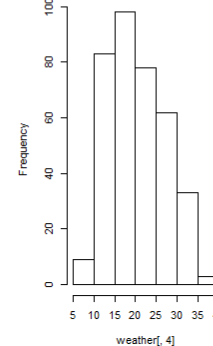
```
hist( weather[ , 6 ] )
```



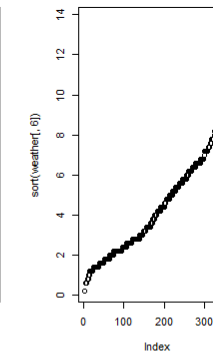
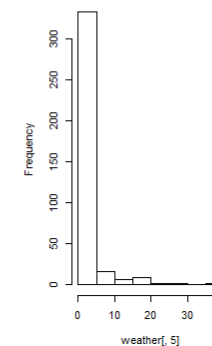
Histogram of weather[, 3]



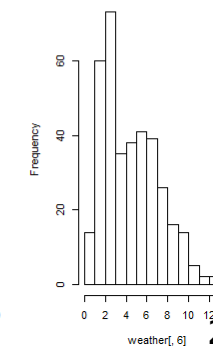
Histogram of weather[, 4]



Histogram of weather[, 5]



Histogram of weather[, 6]



# 分布指標 - weather

# skewness( ), kurtosis( ), in timeDate, 更詳細的分布指標

install.packages( "timeDate" )

library( timeDate )

skewness( weather[ , 3:6] )

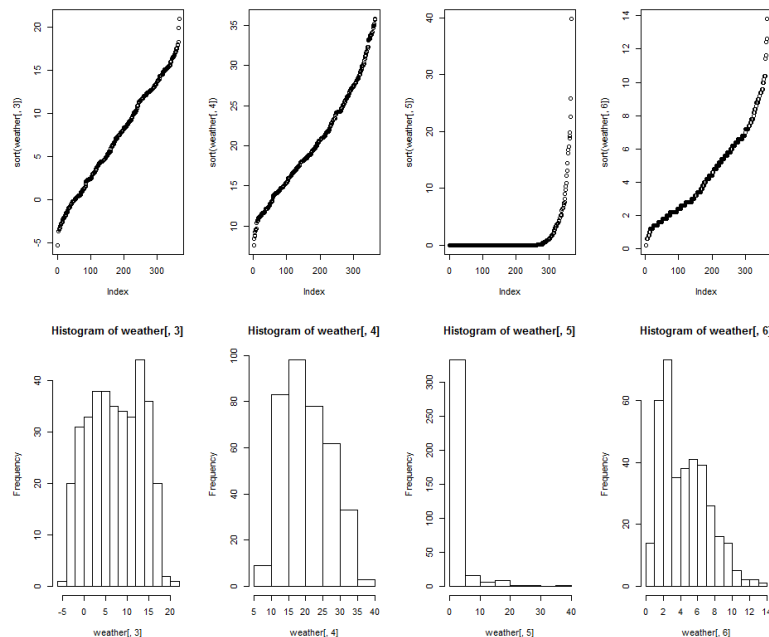
kurtosis( weather [ , 3:6] )

> skewness( weather[ , 3:6] )

MinTemp	MaxTemp	Rainfall	Evaporation
-0.003779725	0.347510625	4.552606775	0.658228261

> kurtosis( weather [ , 3:6] )

MinTemp	MaxTemp	Rainfall	Evaporation
-1.1256902	-0.7636094	26.2397007	-0.2087607



**偏度**：衡量資料的偏倚程度或對稱程度

- = 0: 正態分布，完全對稱
- [-1, 1]: 對稱性較強，不存在左偏或右偏
- > 1: 右偏
- < -1: 左偏

**峰度**：衡量資料的分布型態的陡緩程度，集中或分散

- = 0: 集散程度與正態分布相同，為標準峰度
- > 0: 比正態分布較為陡峭，為尖頂峰度
- < 0: 比正態分布較為平坦，為平頂峰度

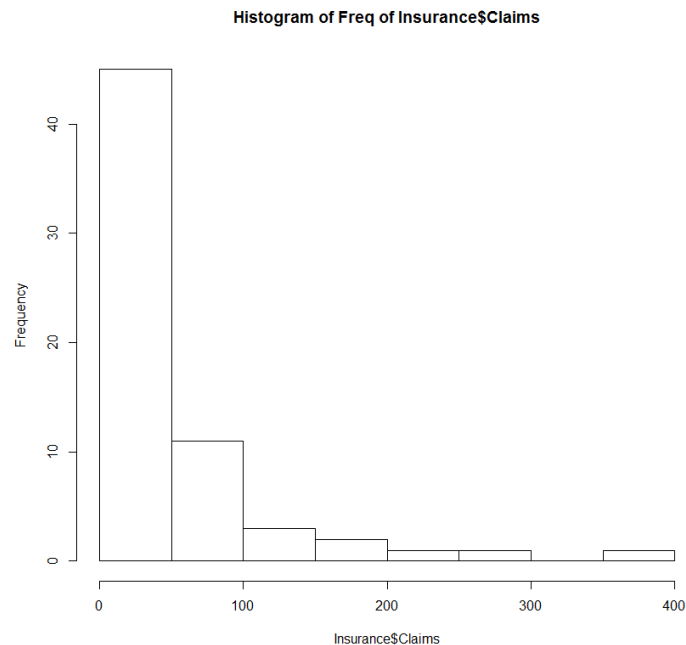
# 視覺化探索 - 長條圖

# 長條圖

- **長條圖**：一種簡單快速探索資料分布的方法
  - 將連續類型的資料分成幾個等間距的組，以矩形的高低來顯示資料的頻數或頻率
  - 有時可同時顯示出資料的密度曲線

```
# install.packages( "Hmisc" )
```

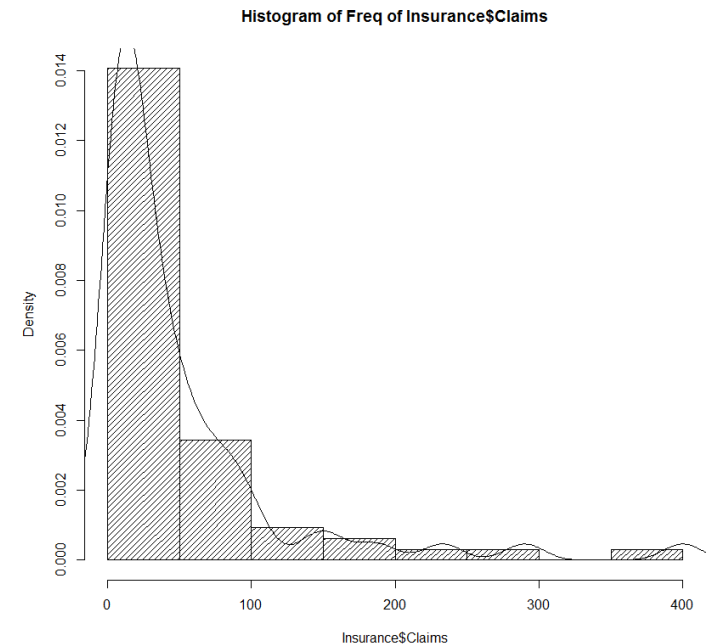
```
hist( Insurance$Claims, main = "Histogram of Freq of Insurance$Claims" )
```



# 長條圖

```
hist( Insurance$Claims, freq = FALSE, density = 20, main = "Histogram of  
Freq of Insurance$Claims" )
```

```
lines( density( Insurance$Claims ) )
```



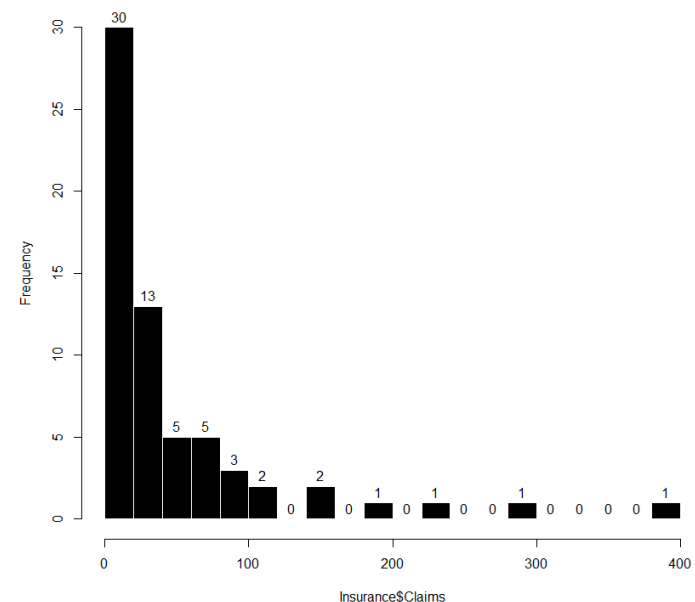


```
hist( Insurance$Claims, breaks = 20, labels = TRUE, col = "black", border =  
"white", main = "Histogram of Freq of Insurance$Claims with 20 bars" )
```

```
out <- hist( Insurance$Claims, breaks = 20, labels = TRUE, col = "black",  
border = "white", main = "Histogram of Freq of Insurance$Claims with 20  
bars" )
```

```
str( out )
```

Histogram of Freq of Insurance\$Claims with 20 bars



List of 6

```
$ breaks : num [1:21] 0 20 40 60 80 100 120 140 160 180 ...  
$ counts : int [1:20] 30 13 5 5 3 2 0 2 0 1 ...  
$ density : num [1:20] 0.02344 0.01016 0.00391 0.00391 0.00234 ...  
$ mids : num [1:20] 10 30 50 70 90 110 130 150 170 190 ...  
$ xname : chr "Insurance$Claims"  
$ equidist: logi TRUE
```

```
- attr(*, "class")= chr "histogram"
```

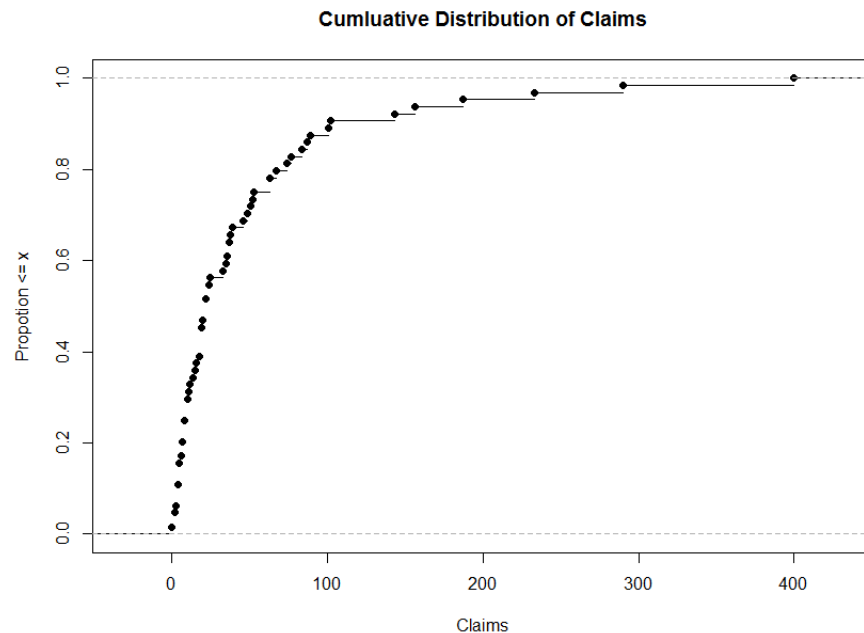
# 視覺化探索 - 累積分布圖

# 累積分布圖

- 累積分布圖：
  - 可以觀察資料分布情形

```
dt <- ecdf( Insurance$Claims )
```

```
plot( dt, xlab = "Claims", ylab = "Propotion <= x", main = "Cumluative Distribution of Claims" )
```

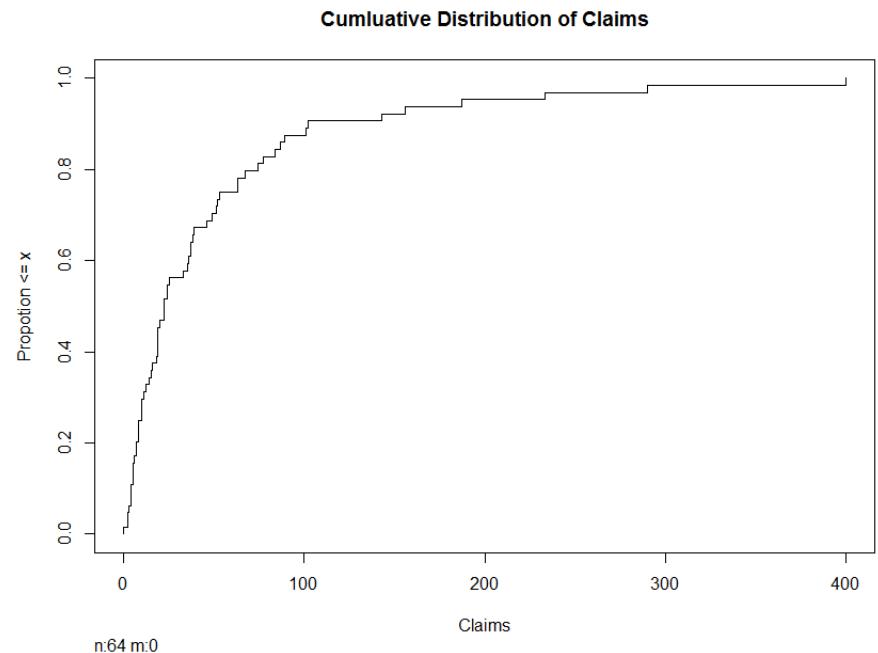


# 累積分布圖

```
install.packages("Hmisc")
```

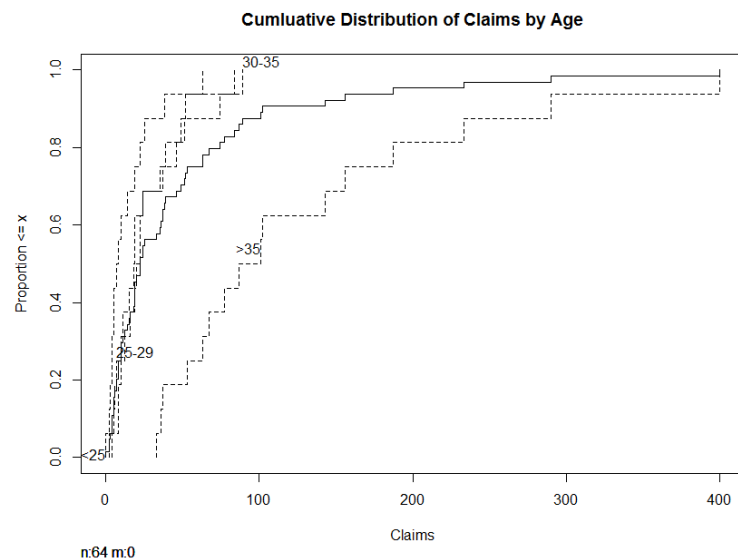
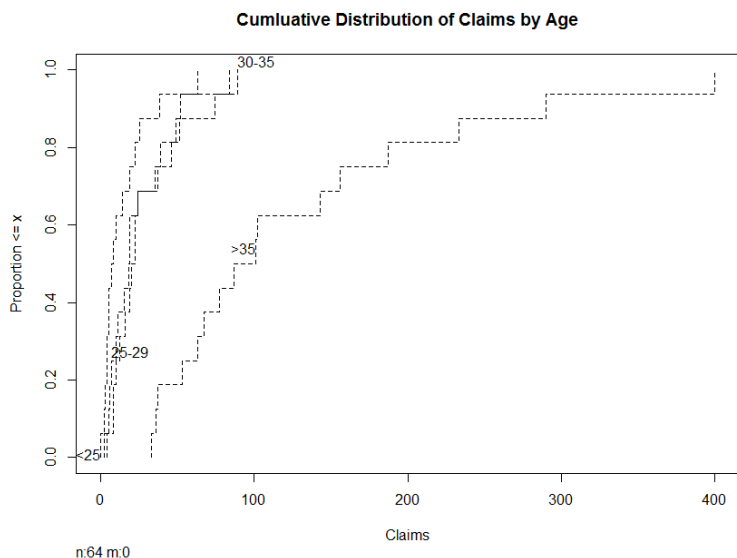
```
library(Hmisc)
```

```
Ecdf( Insurance$Claims, xlab = "Claims", ylab = "Propotion <= x", main =  
"Cumluative Distribution of Claims" )
```



# 累積分布圖

```
data_plot <- with( Insurance, rbind(
data.frame( var1 = Claims[ Age == "<25" ], var2 = "<25" ),
data.frame( var1 = Claims[ Age == "25-29" ], var2 = "25-29" ),
data.frame( var1 = Claims[ Age == "30-35" ], var2 = "30-35" ),
data.frame( var1 = Claims[ Age == ">35" ], var2 = ">35" ) ) )
Ecdf( data_plot$var1, group = data_plot$var2, lty = 2, label.curves=1:4,
xlab = "Claims", main = "Cumulative Distribution of Claims by Age" )
Ecdf( Insurance$Claims, add = TRUE )
```



## 視覺化探索 - 箱形圖

# 箱形圖 or 盒鬚圖

- 箱形圖 or 盒鬚圖：

- 可以較深入展現資料分布情況，列出重要分位點，將異數剝離出來

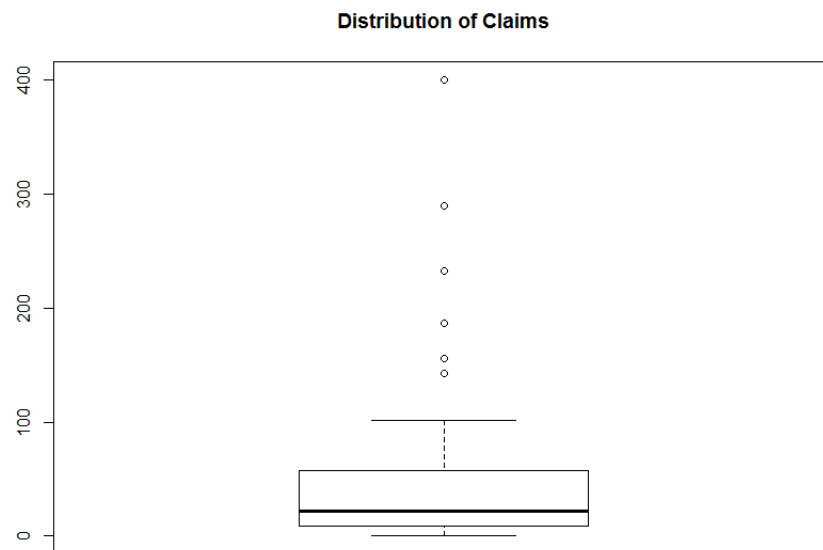
```
Claims_bp <- boxplot( Insurance$Claims, main = "Distribution of Claims" )
```

```
Claims_bp$stats
```

```
> Claims_bp$stats
```

```
      [,1]
[1,]    0
[2,]    9
[3,]   22
[4,]   58
[5,]  102
attr(,"class")
```

```
"integer"
```



# 箱形圖 or 盒鬚圖

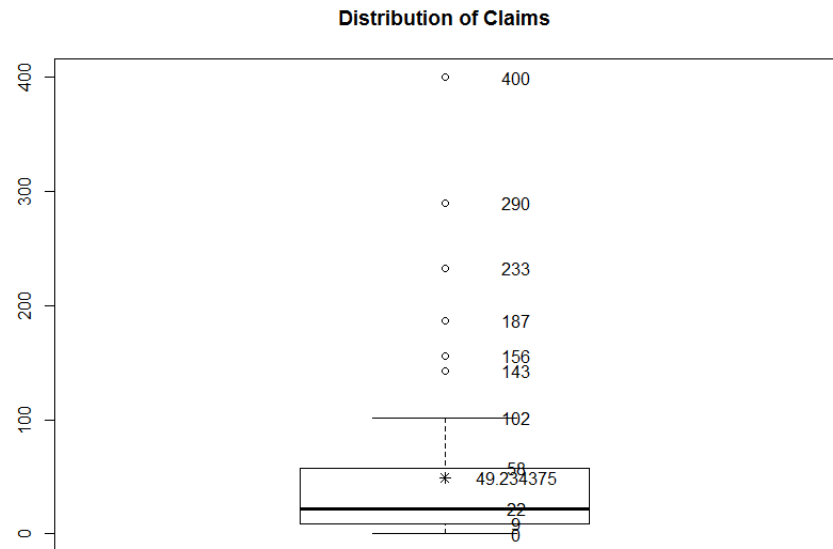
- 標記資料點

```
points( x=1, y = mean( Insurance$Claims ), pch = 8 )
```

```
Claims_points <- as.matrix( Insurance$Claims[ which( Insurance$Claims > 102 ) ], 6, 1 )
```

```
Claims_text <- rbind( Claims_bp$stats, mean( Insurance$Claims), Claims_points )
```

```
for( i in 1:length( Claims_text ) ) text( x = 1.1, y = Claims_text[ i, ], labels = Claims_text[ i, ] )
```

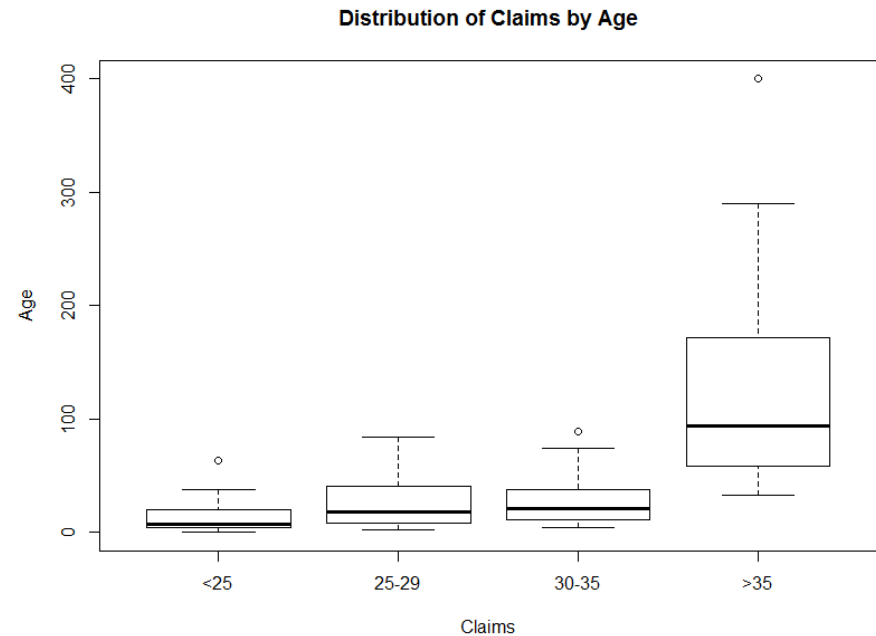
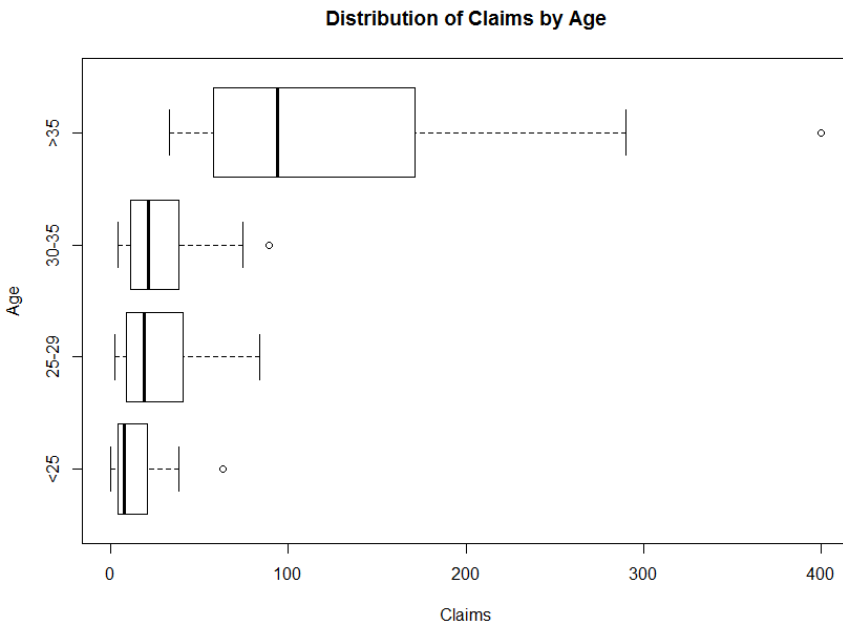




# 箱形圖 or 盒鬚圖

- 一張圖容納多個箱形圖

```
boxplot( var1~var2, data = data_plot, horizontal = TRUE, main =  
"Distribution of Claims by Age", xlab = "Claims", ylab = "Age" )
```



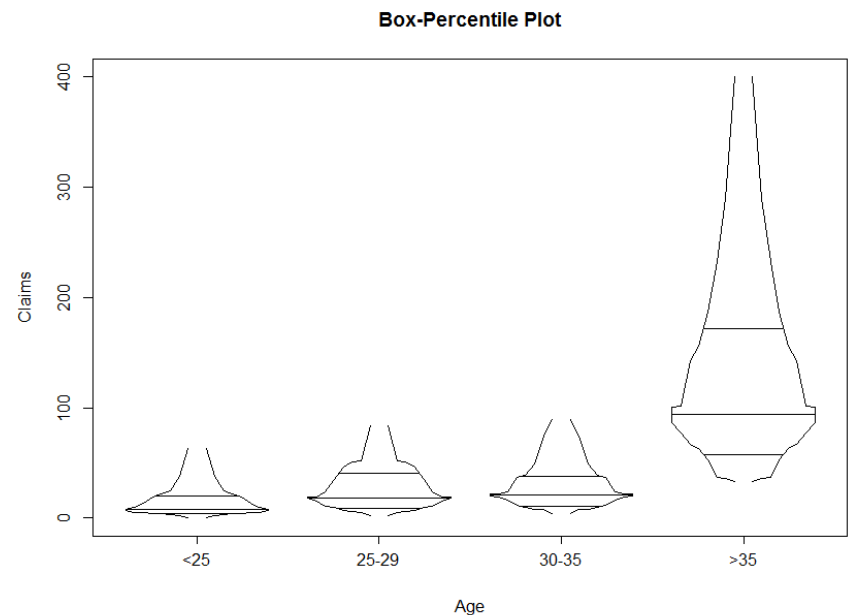
# 箱形圖 or 盒鬚圖

## ■ 比例箱形圖

```
data_bp <- list(
```

```
data_plot$var1[ which( data_plot$var2 == "<25" ) ],  
data_plot$var1[ which( data_plot$var2 == "25-29" ) ],  
data_plot$var1[ which( data_plot$var2 == "30-35" ) ],  
data_plot$var1[ which( data_plot$var2 == ">35" ) ] )
```

```
bpplot( data_bp, name = c( "<25", "25-29", "30-35", ">35" ), ylab =  
"Claims", xlab = "Age" )
```



## 視覺化探索 - 橫條圖

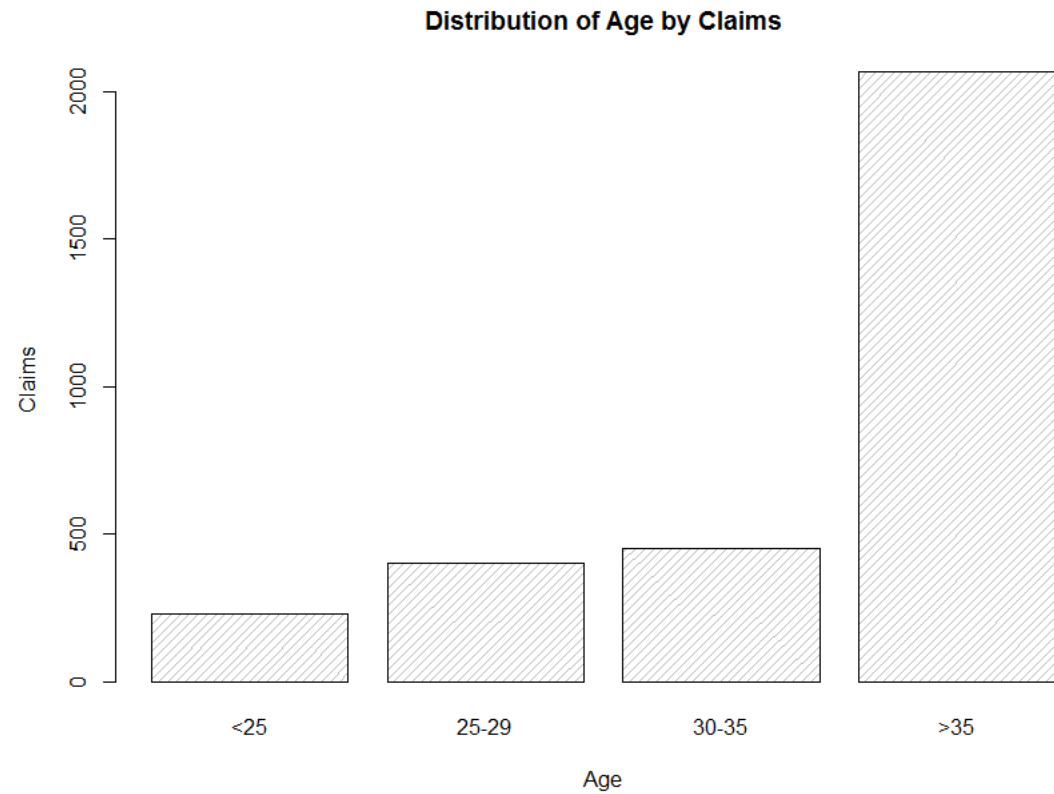
# 橫條圖

- 橫條圖：
  - 主要針對離散型變數，每一個水準自然成為一個條形來顯示該水準的設定值

```
Claims_Age <- with( Insurance,  
c( sum( Claims[ which( Age == "<25" ) ] ),  
sum( Claims[ which( Age == "25-29" ) ] ),  
sum( Claims[ which( Age == "30-35" ) ] ),  
sum( Claims[ which( Age == ">35" ) ] ) ) )
```

```
barplot( Claims_Age, names.arg = c( "<25", "25-29", "30-35", ">35" ),  
density = rep( 20, 4), main = "Distribution of Age by Claims", ylab =  
"Claims", xlab = "Age" )
```

# 橫條圖



- 分組 橫條圖：

```
 Holders_Age <- with( Insurance,  
  c( sum( Holders[ which( Age == "<25" ) ] ),  
    sum( Holders[ which( Age == "25-29" ) ] ),  
    sum( Holders[ which( Age == "30-35" ) ] ),  
    sum( Holders[ which( Age == ">35" ) ] ) ) )
```

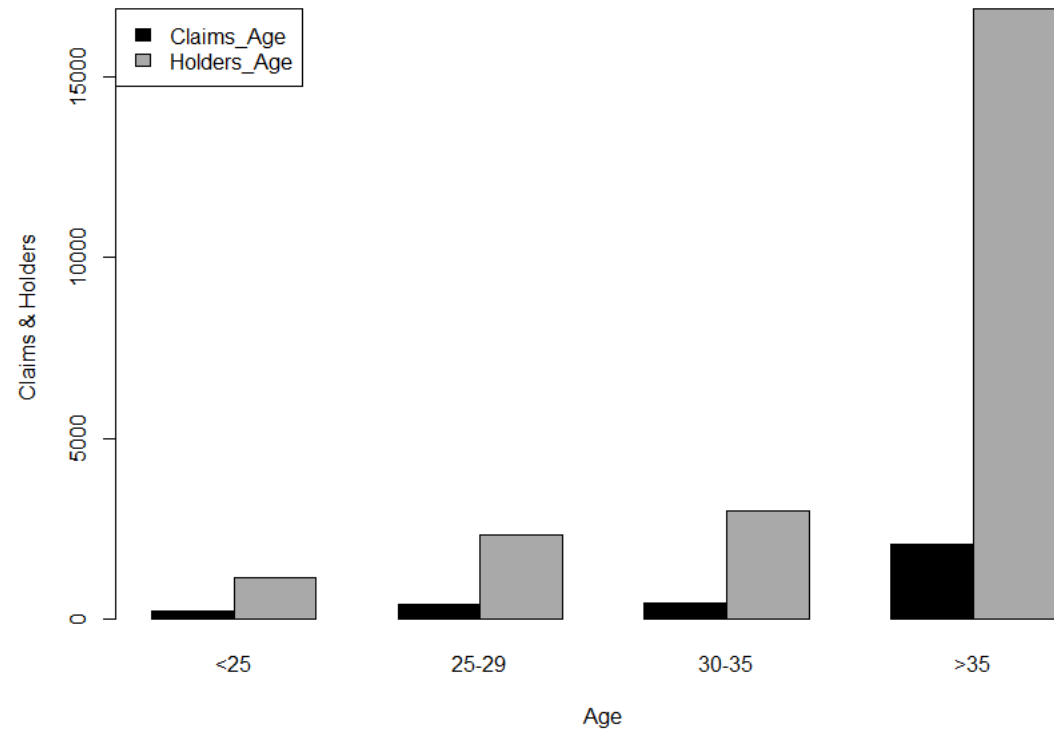
```
data_bar <- rbind( Claims_Age, Holders_Age )
```

```
barplot( data_bar, names.arg = c( "<25", "25-29", "30-35", ">35" ), beside  
= TRUE, density = rep( 20, 4), main = "Age Distribution by Claims and  
Holders", ylab = "Claims & Holders", xlab = "Age", col = c( "black",  
"darkgrey" ) )
```

```
legend( x = "topleft", rownames( data_bar ), fill = c( "black", "darkgrey" ) )
```

# 橫條圖

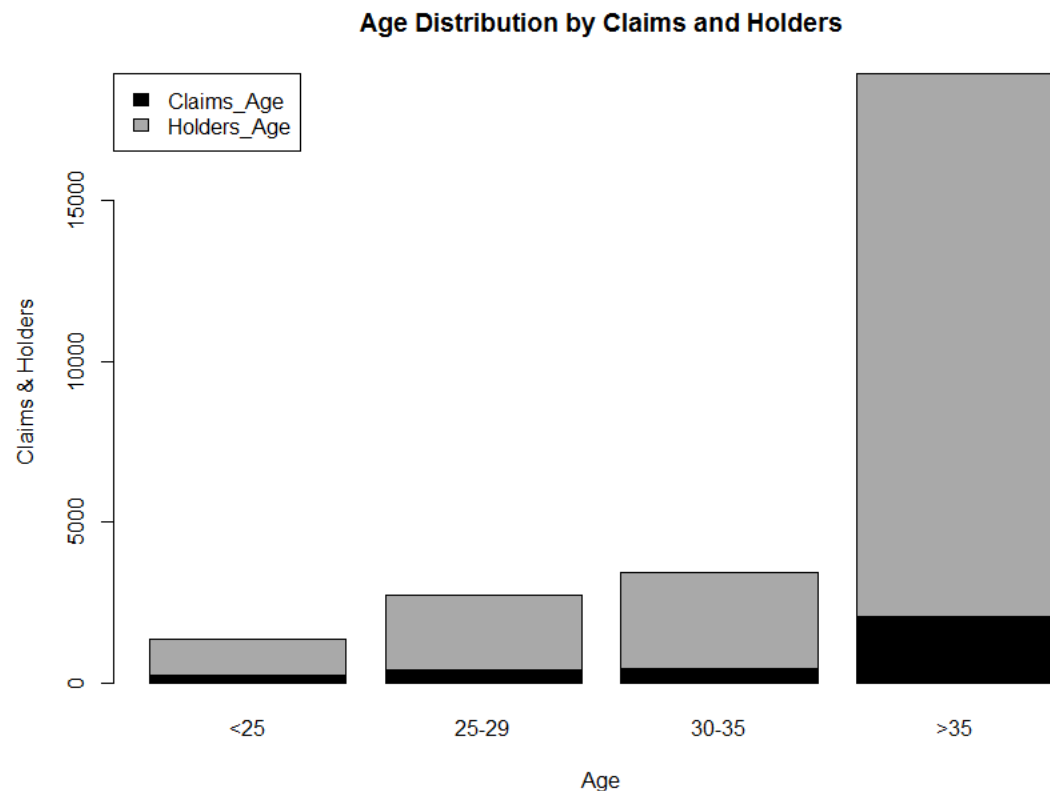
Age Distribution by Claims and Holders



## ■ 堆疊 橫條圖：

```
barplot( data_bar, names.arg = c( "<25", "25-29", "30-35", ">35" ), main = "Age Distribution by Claims and Holders", ylab = "Claims & Holders", xlab = "Age", col = c( "black", "darkgrey" ) )
```

```
legend( x = "topleft", rownames( data_bar ), fill = c( "black", "darkgrey" ) )
```





# 視覺化探索 - 點陣圖

# 點陣圖

- 點陣圖：
  - 用於呈現離散型變數各取樣水準的分布情形，用點與背景格線代替線條

```
dotchart( data_bar, main = "Age Distribution by Claims and Holders", xlab =  
"Claims & Holders", pch = 1:2 )
```

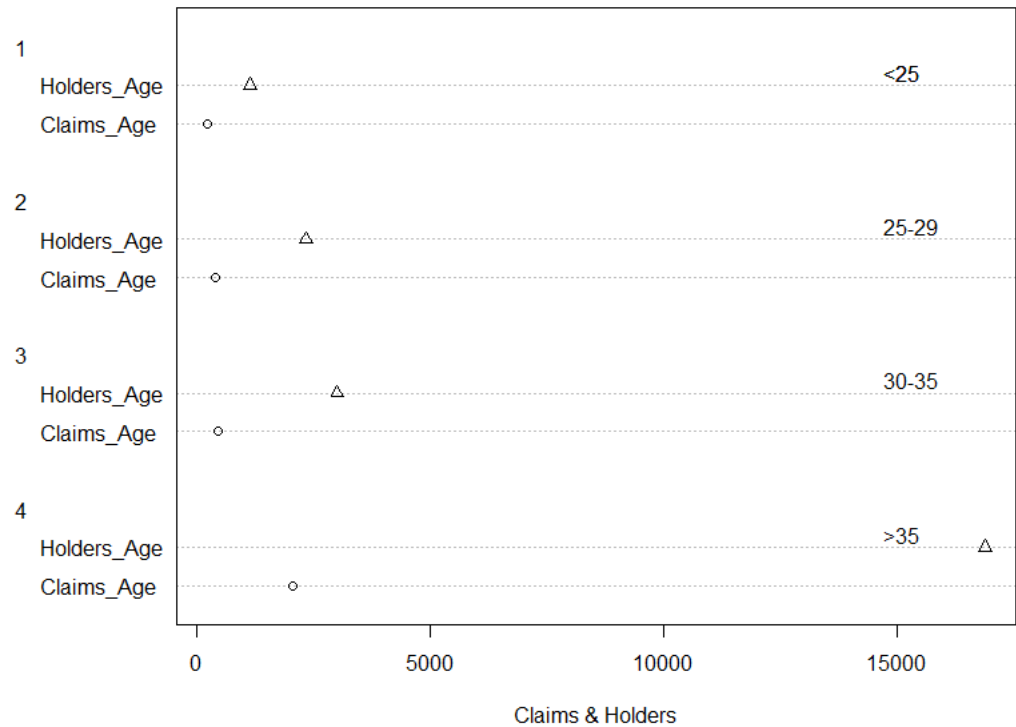
```
legend( x = 14000, y = 15, "<25", bty = "n" )
```

```
legend( x = 14000, y = 11, "25-29", bty = "n" )
```

```
legend( x = 14000, y = 7, "30-35", bty = "n" )
```

```
legend( x = 14000, y = 3, ">35", bty = "n" )
```

Age Distribution by Claims and Holders

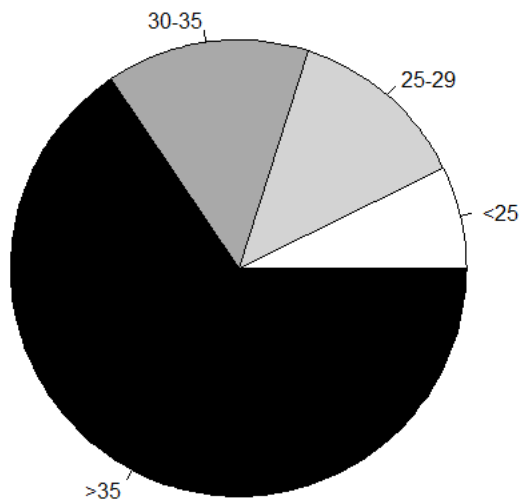


# 視覺化探索 - 圓形圖

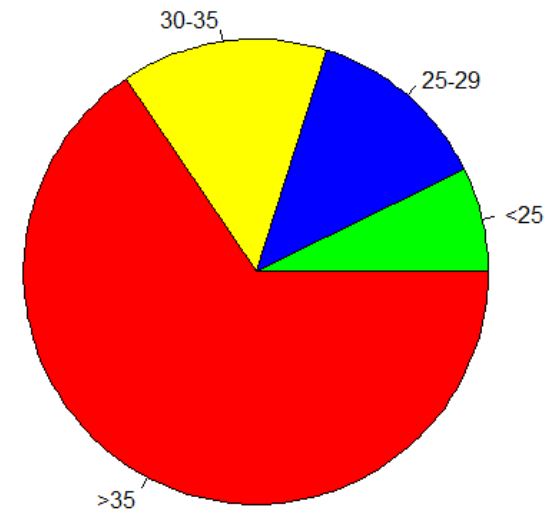
# 圓形圖

```
pie( Claims_Age, labels = c( "<25", "25-29", "30-35", ">35" ), main = "Pie  
Chart of Age by Claims", col = c( "white", "lightgrey", "darkgrey",  
"black" ) )
```

Pie Chart of Age by Claims



Pie Chart of Age by Claims



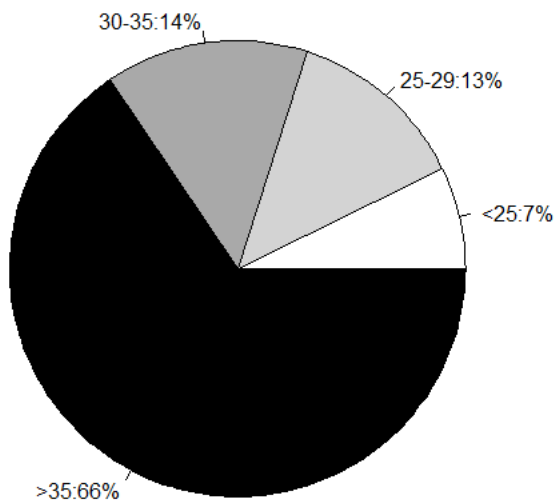
# 圓形圖

```
percent <- round( Claims_Age / sum( Claims_Age ) * 100 )
```

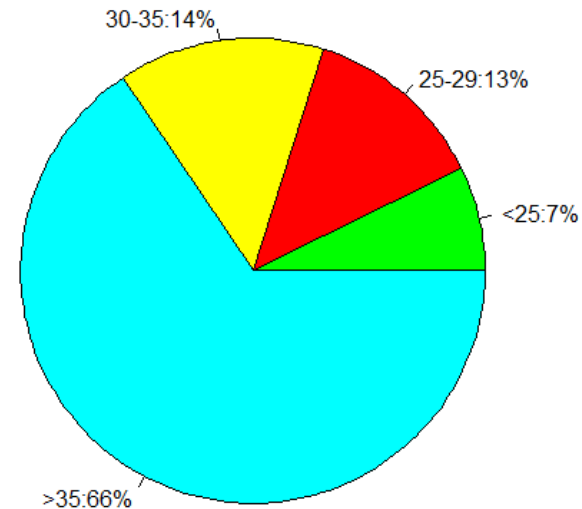
```
label <- paste( c( "<25", "25-29", "30-35", ">35" ), ":", percent, "%",  
sep="" )
```

```
pie( Claims_Age, labels = label, main = "Pie Chart of Age by Claims", col =  
c( "white", "lightgrey", "darkgrey", "black" ) )
```

Pie Chart of Age by Claims



Pie Chart of Age by Claims



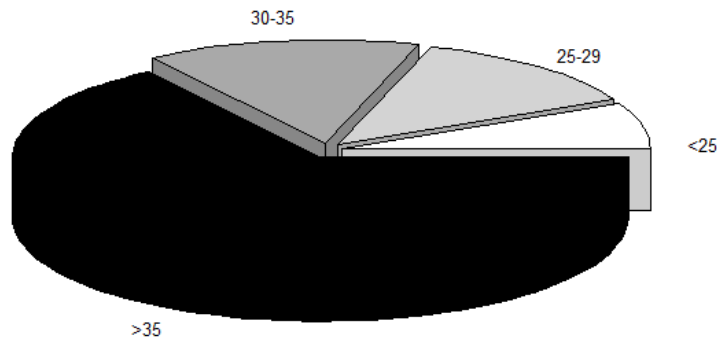
# 圓形圖 - 3D

```
install.packages( "plotrix" )
```

```
library( plotrix )
```

```
pie3D( Claims_Age, labels = c( "<25", "25-29", "30-35", ">35" ), explode  
= 0.05, main = "3D Pie Chart of Age by Claims", labelcex = 0.8, col =  
c( "white", "lightgrey", "darkgrey", "black" ) )
```

3D Pie Chart of Age by Claims



3D Pie Chart of Age by Claims

