資料視覺化

國立宜蘭大學資訊工程系
吳政瑋 助理教授
wucw@niu.edu.tw

教學目標

- 了解資料視覺化的目的
- 了解常見的資料視覺化圖表
- 了解如何以軟體及工具製作圖表

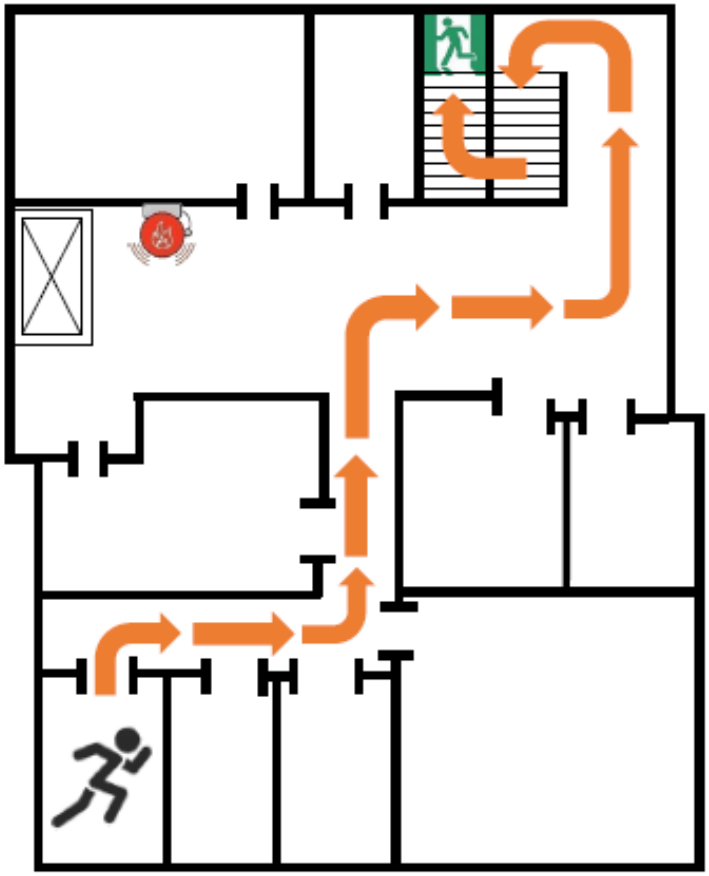
眼力大考驗

假設發生火警，在煙霧彌漫中，你衝到門口，在那裡你看到緊急出口標語牌...

離開辦公室，右轉，有一條300公尺長的走廊直走到底，你會看到一間大會議室。向左轉，還有一條360公尺的走廊直走到底，在你的左手邊有一個火警警報器，旁邊是電梯；右手邊有一座樓梯。不要走到電梯；右轉，從另一條360公尺長的走廊直走到底，左轉進入樓梯間。走下兩階樓梯後，從樓梯底部的門離開大樓。

眼力大考驗

假設發生火警，在煙霧彌漫中，你衝到門口，在那裡你看到緊急出口標語牌...



資料視覺化的目的

- **資料視覺化(Data Visualization)**是要用圖形或是表格的方式來呈現資料。一個成功的視覺化圖表，能夠清楚的呈現資料的特性，以及資料間或是屬性值的關係，而且可以輕易的讓人看圖示意。
- 在日常生活中，視覺化的圖表可以用來解釋氣象、經濟及**選舉**的預測結果。
- 有時會將**資料探勘(Data Mining)**的視覺化技術稱為**視覺化資料探勘(Visual Data Mining)**。

敘述性統計(Descriptive Statistics)

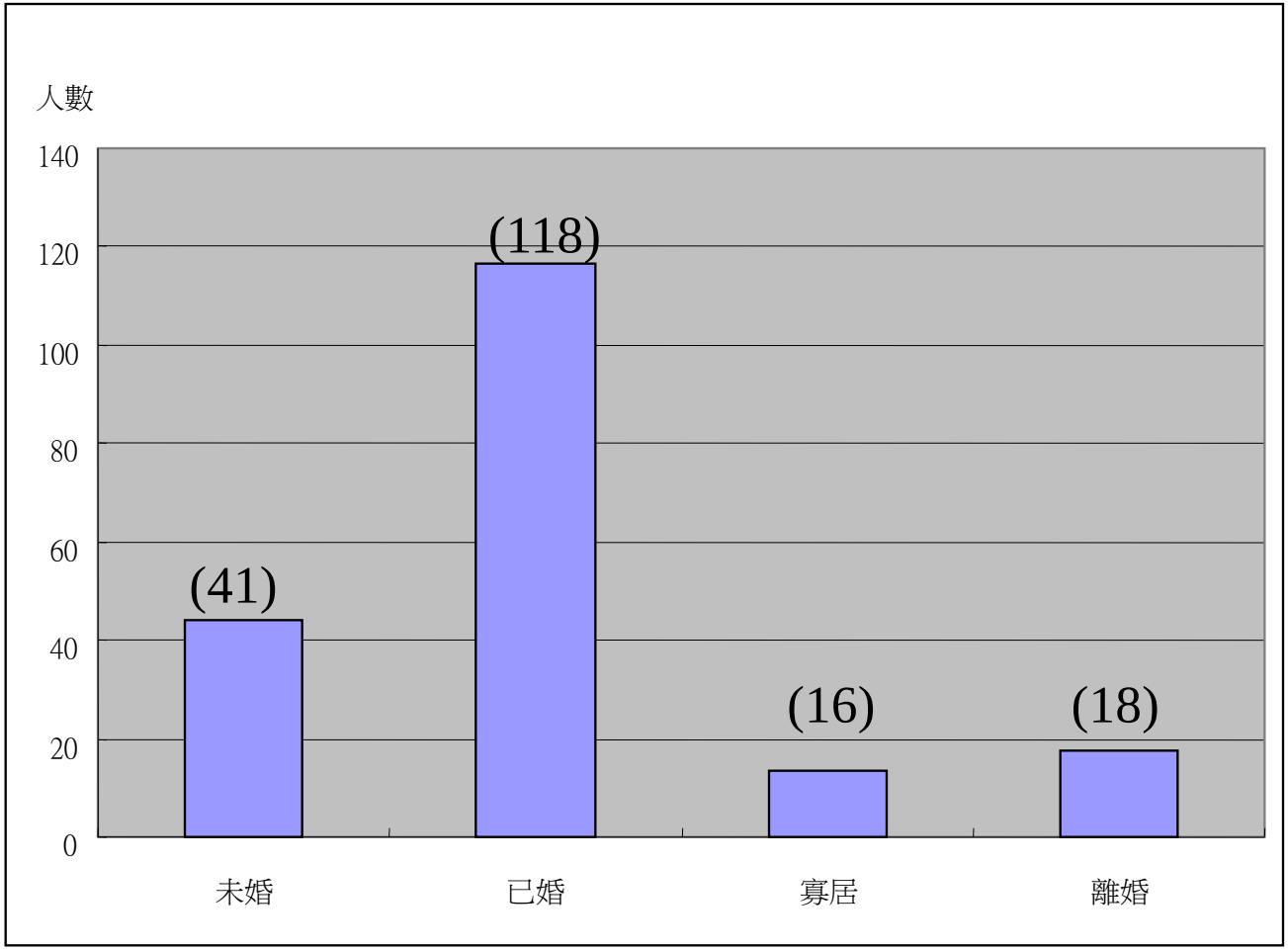
- **敘述性統計(Descriptive Statistics)**在**資料探勘**領域中，又稱為**探索式資料分析 (Exploratory Data Analysis，簡稱EDA)**。
- EDA是由有統計界畢卡索之稱的Tukey所提出，主要概念是透過敘述性統計、統計繪圖、視覺化等快速簡易的方式，從各種面向先了解資料的狀況，以利後續分析。

EDA常見的資料圖表類型

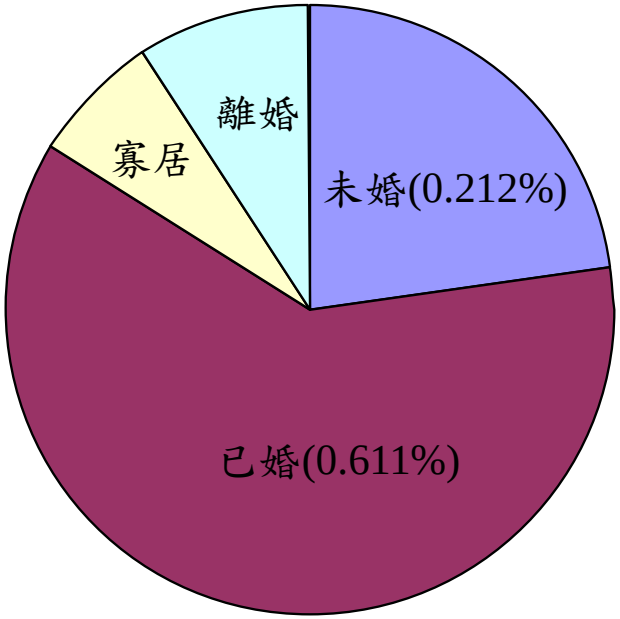
- 類別屬性
  - 長條圖(Bar Chart)
  - 圓形圖(Pie Chart)

- 數值屬性
  - 莖葉圖(Stem Plot / Stem-and-leaf plot)
  - 直方圖(Histogram)
  - 盒型圖(Box Plot)
  - 散佈圖(Scatter Plot)
  - 折線圖(Line Chart)

長條圖(Bar Chart)



圓形圖(Pie Chart)



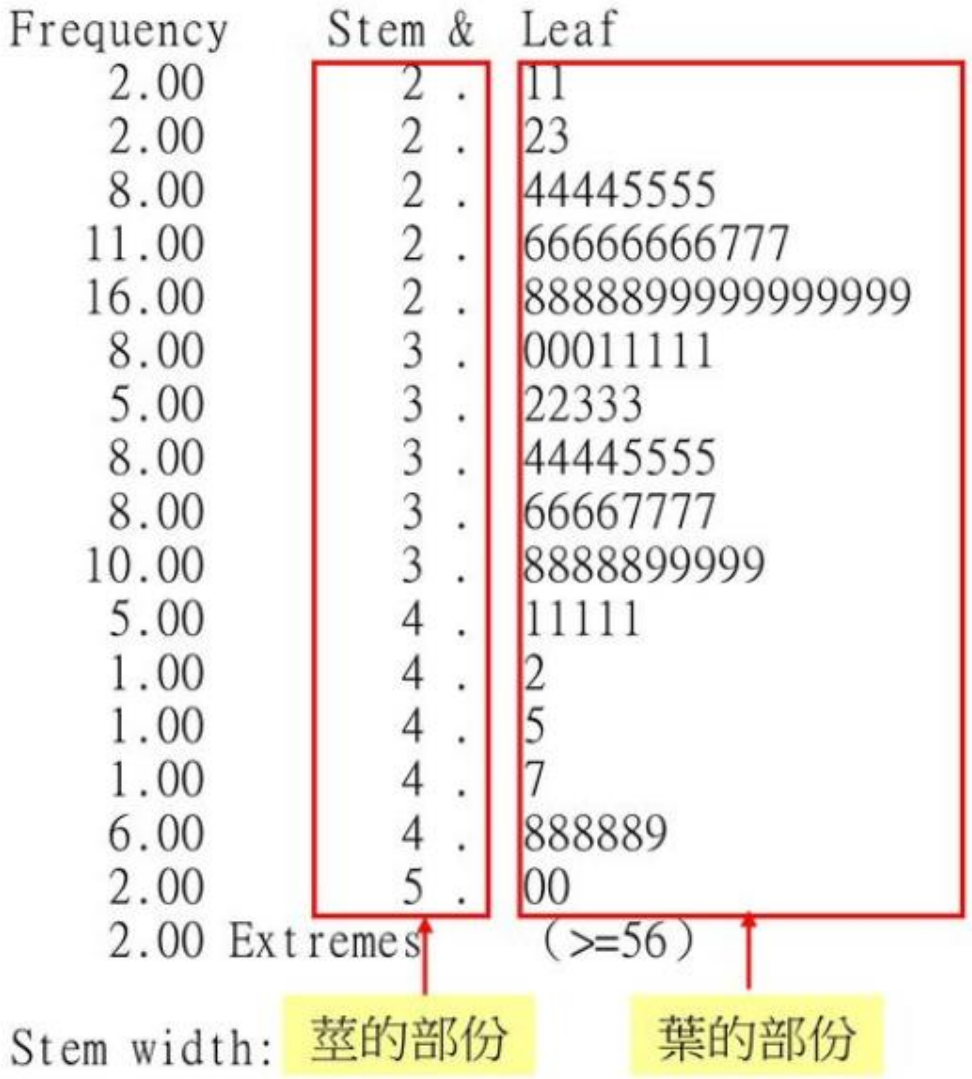
注意事項:

1. 使用 Pie chart 可顯示每一類所佔的百分比
2. Bar chart 可作到相同的效果
3. 換算成百分比的過程可能產生誤差

莖葉圖(Stem-and-Leaf Plot)

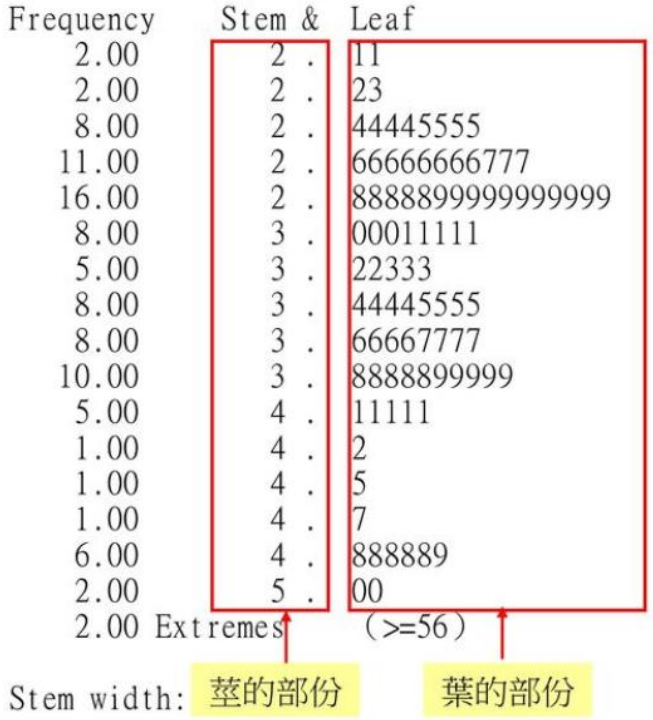
- 普林斯頓大學John Tukey教授於1977年所發展。
- **莖葉圖**最適合**二位數資料**之呈現，如：考試成績。
- **莖葉圖**的製作是將每個觀察值切割為**莖**及**葉**兩部份，中間以垂直線隔開。**莖**為觀察值中間**十位數或以上之數字**；**葉**則為**個位數**的數字。

莖葉圖(Stem-and-Leaf Plot)

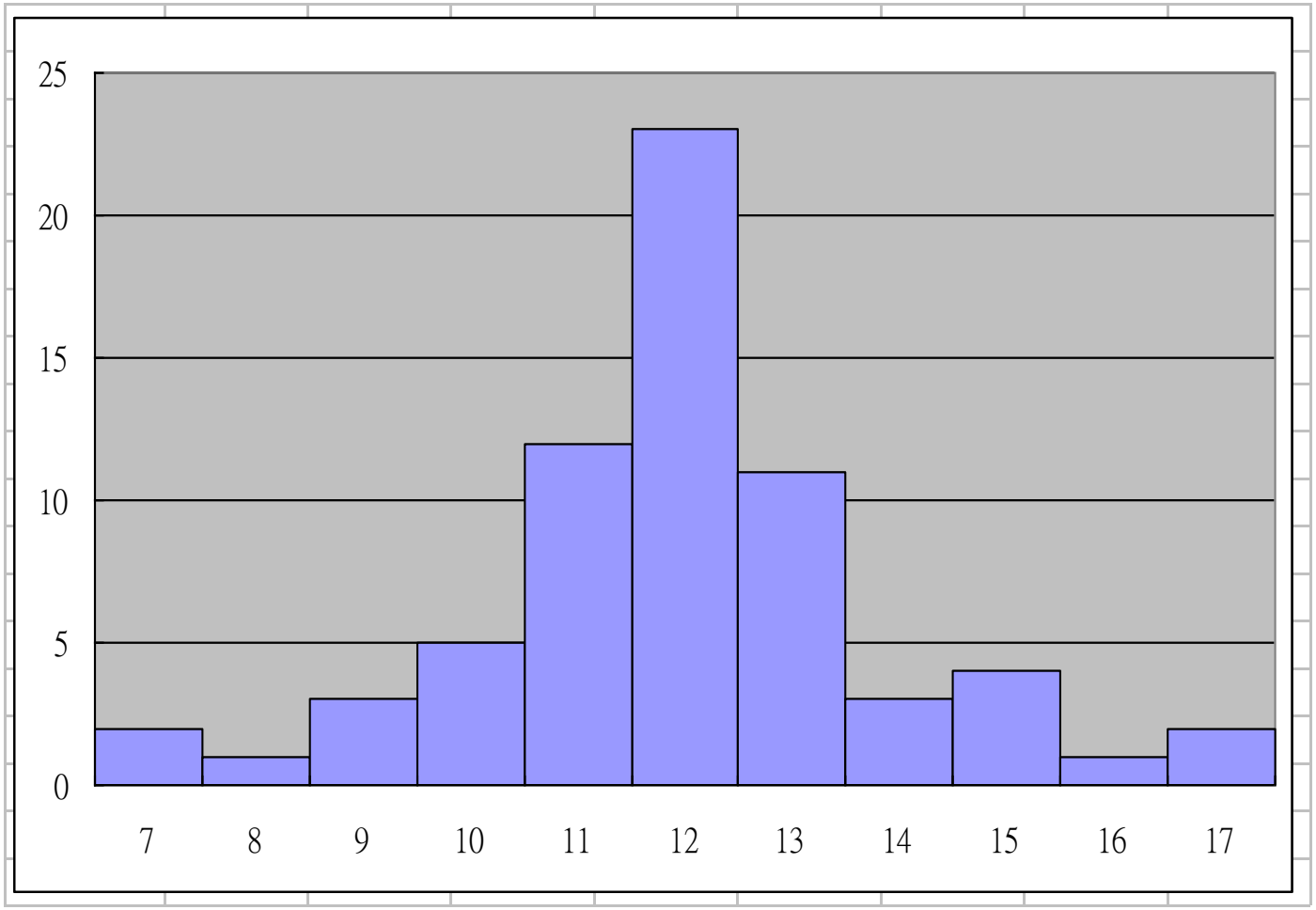


莖葉圖(Stem-and-Leaf Plot)

- **莖葉圖**的主要優點為兼具**次數分配表**與**長條圖**之雙重優點。
- 莖葉圖**不會失去原資料值**，所有原資料值在莖葉圖上一目瞭然。

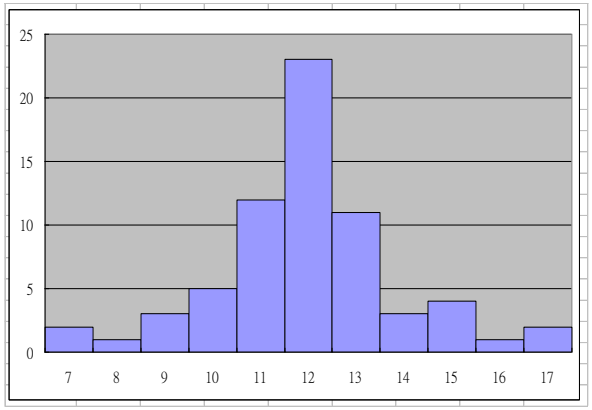


直方圖(Histogram)

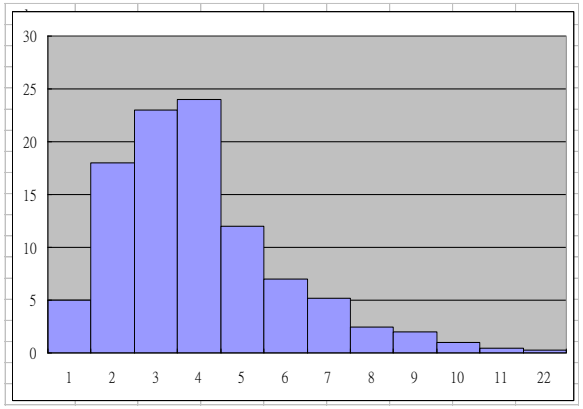


直方圖
對稱分佈與偏斜分佈

- **對稱分佈(Symmetric Distribution)**
  - 直方圖的右邊大約為左邊的鏡像

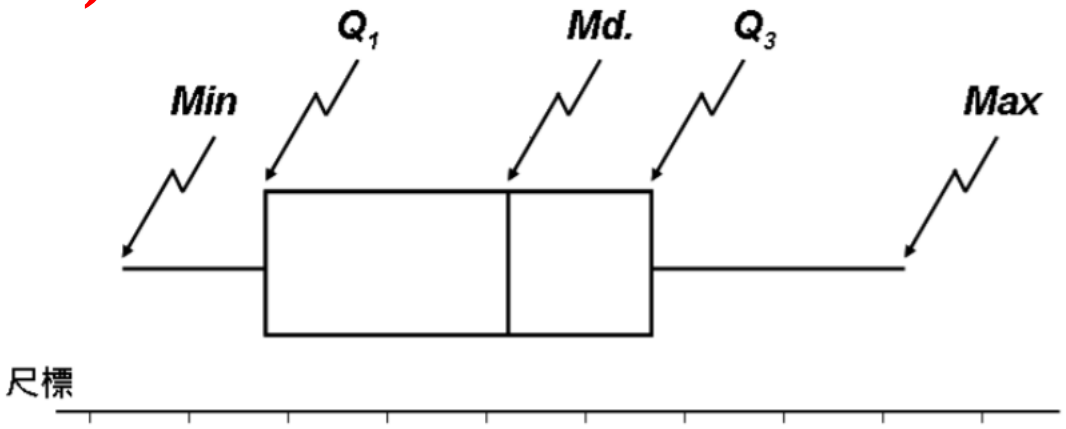


- **偏斜分佈**
  - 右偏(Skew to the right): 右邊延展比左邊遠
  - 左偏(Skew to the left)



箱型圖(Box Plot)

- **箱型圖(Box Plot)**，又稱**盒型圖**、**盒鬚圖**、**盒式圖**、**盒狀圖**或**箱線圖**。
- **箱形圖**於1977年由美國著名統計學家 John Tukey 發明。它為**五數綜合**的圖形化表示法，能顯示出一組數據的**最大值**、**最小值**、**中位數**、及**上下四分位數**，其可用來了解資料的分佈情況，並判斷是否存在**離群值(Outlier)**。



全距(Range)與四分位數(Quartiles) (Cont. 1/2)

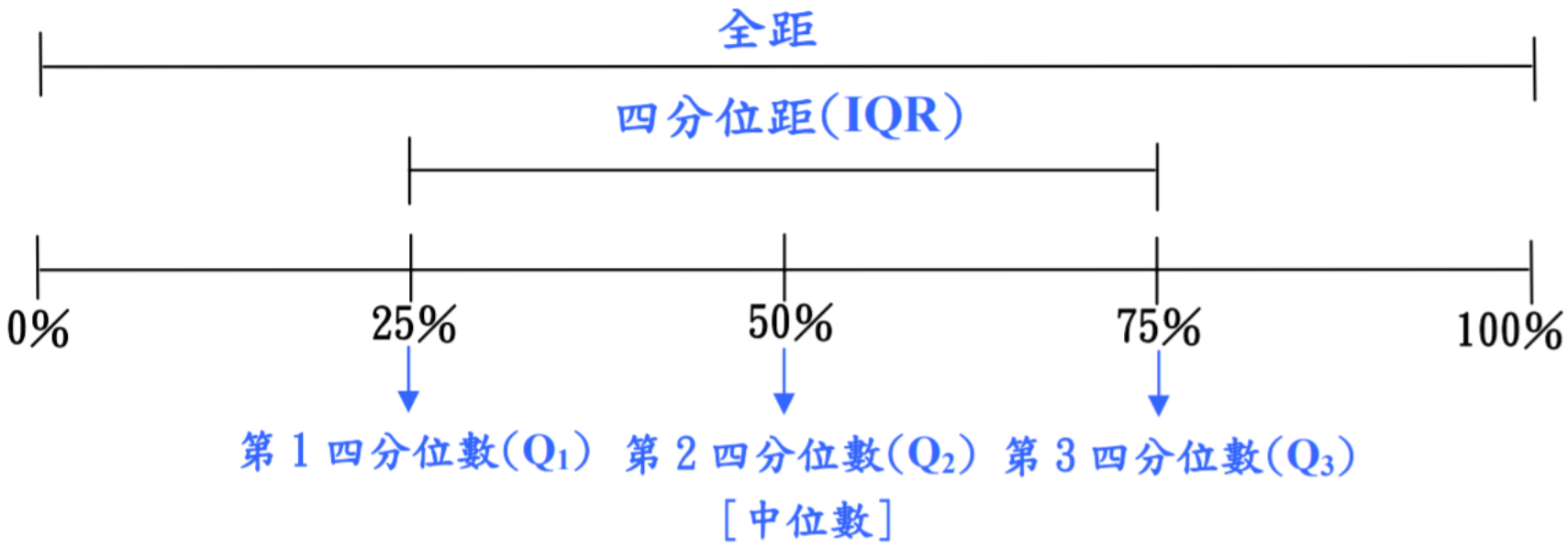
- **全距(Range)**，即一組觀測值的**最大值減去最小值**。
- **第一四分位數**(Q_1)，又稱「**上四分位數**」，等於該樣本中所有數值由小到大排列後**第25%**的數字。
- **第二四分位數**(Q_2)，又稱「**中位數**」，等於該樣本中所有數值由小到大排列後**第50%**的數字。
- **第三四分位數**(Q_3)，又稱「**下四分位數**」，等於該樣本中所有數值由小到大排列後**第75%**的數字。

四分位距與四分位差

- **四分位距(Interquartile-range；IQR)**
  - 計算方式：Q3– Q1

- **四分位差(Quartile Deviation；QD)**
  - 計算方式： (Q3 - Q1)/2

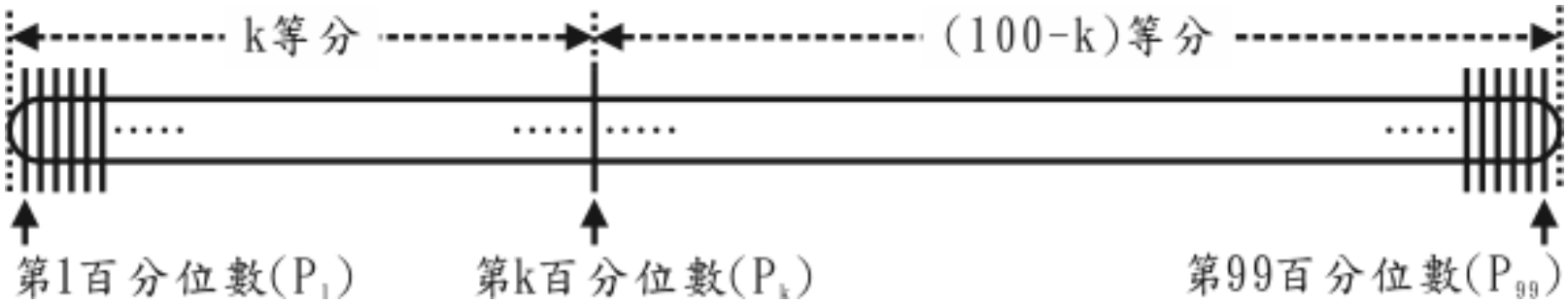
全距、四分位數、四分位距的視覺化圖形



百分位數(Percentile)

- **第 *p* 百分位數(The *p*-th Percentile)**
  - 有 *p*% 的資料，其值低於 *p*-th percentile

- 舉例說明
  - 若P_{30}等於60，則資料組中有30%的資料低於60分。

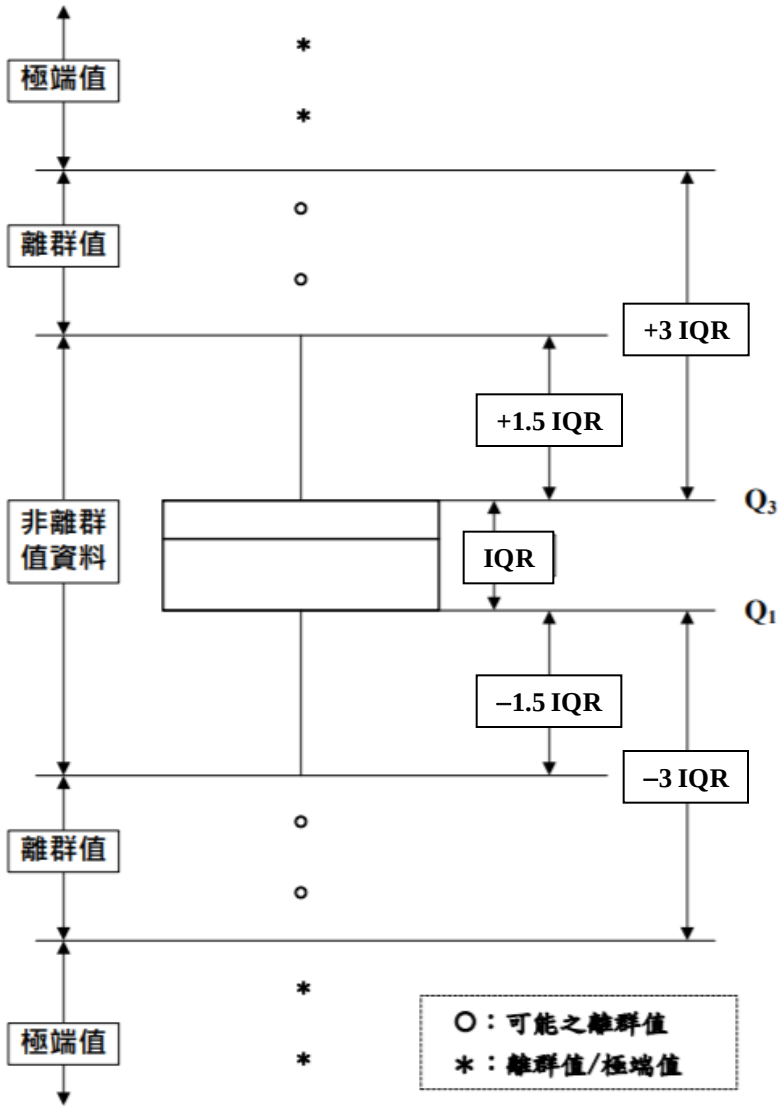


五數綜合(Five-Number Summary)

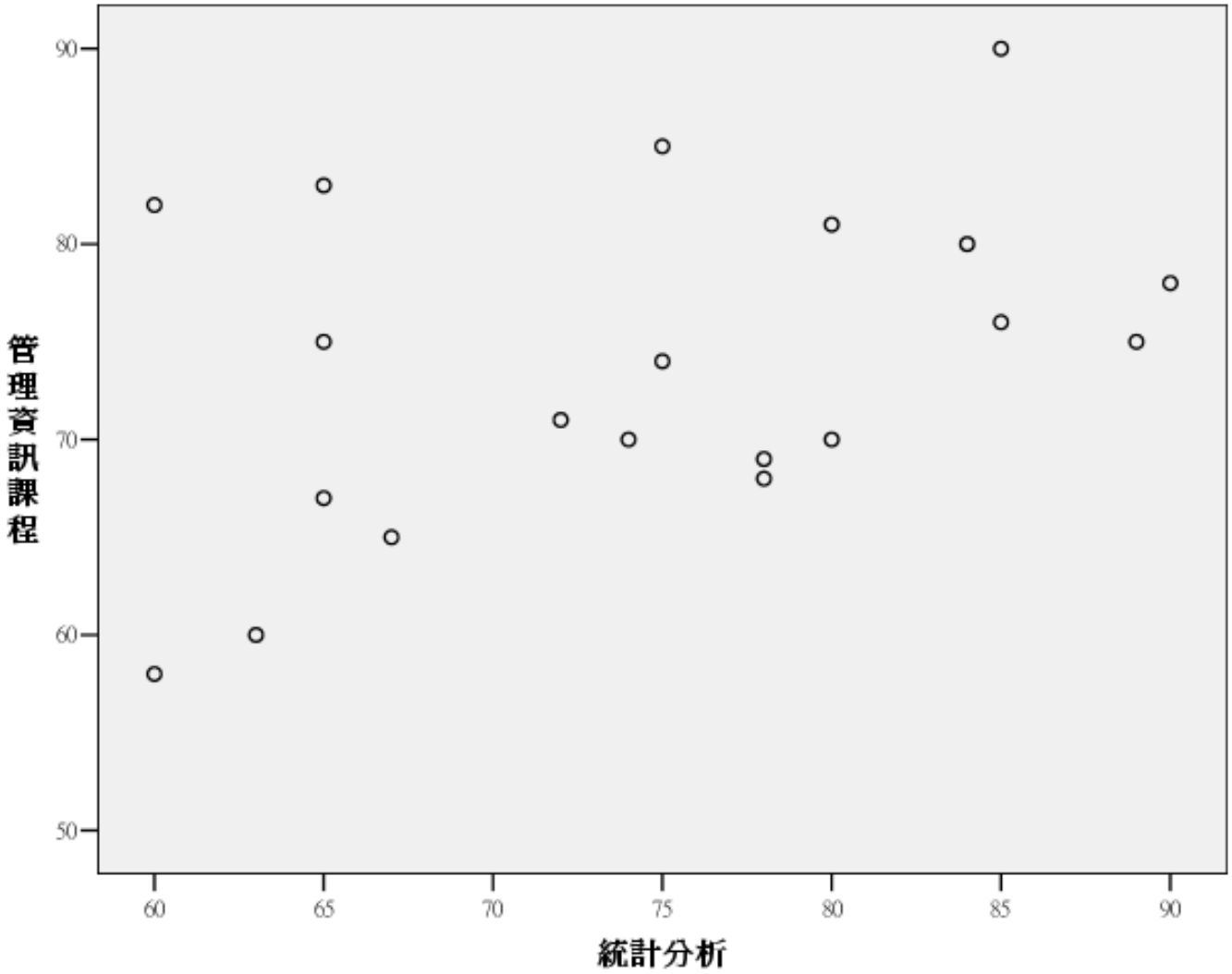
- 五數綜合
  - 最大值(Maximum)
  - 第一四分位數(First Quartile)
  - 中位數(Median)
  - 第三四分位數(Third Quartile)
  - 最小值(Minimum)

箱型圖上的離群值與極端值

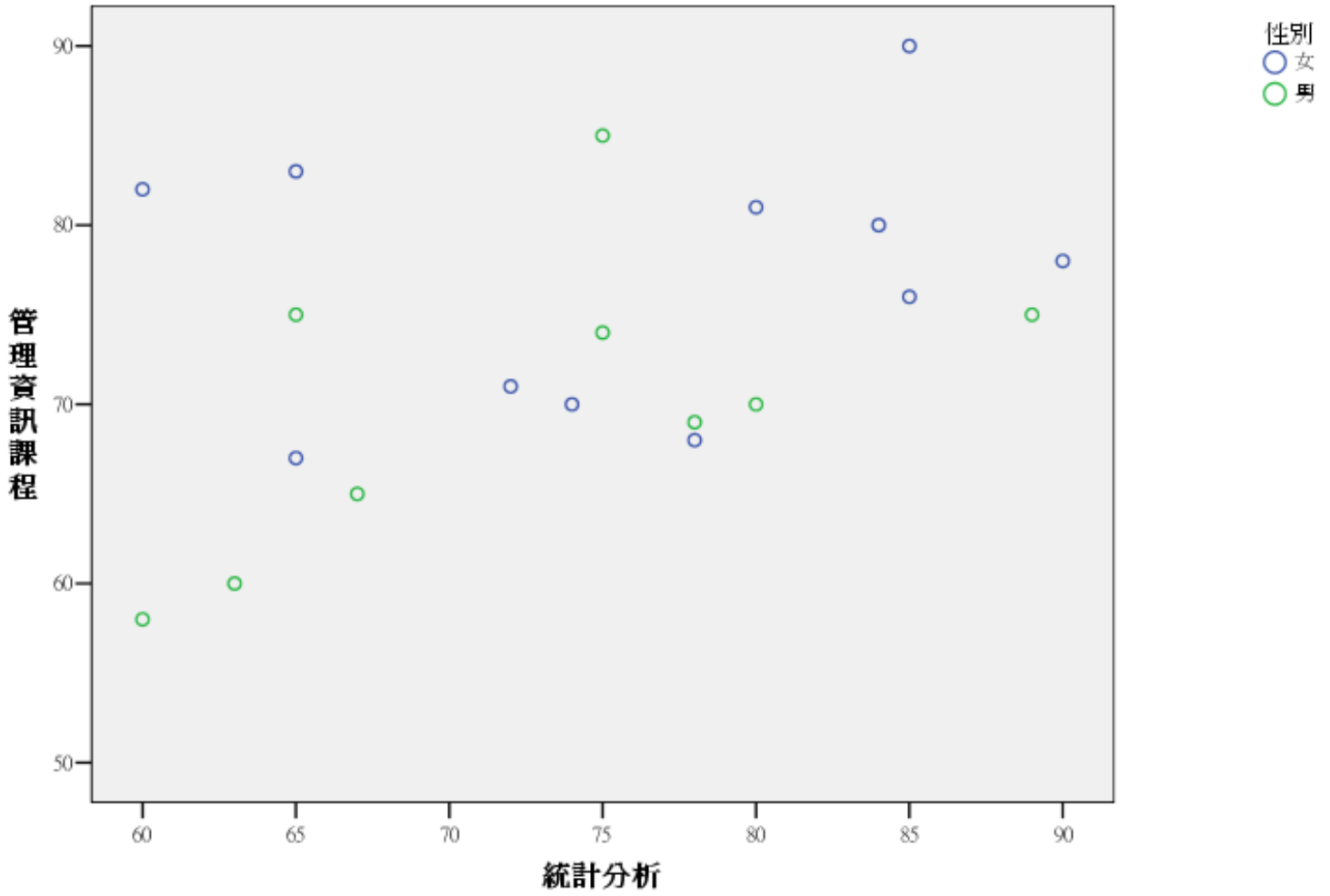
- **離群值(Outlier)**
  - (Q1 – 1.5 × IQR) ~ (Q1 – 3×IQR)
  - (Q3 + 1.5 × IQR) ~ (Q3 + 3×IQR)
- **極端值(Extreme Outlier)**
  - 值小於 Q1 – 3×IQR 或
  - 值大於 Q3 + 3×IQR



散佈圖(Scatter Plot)



條件散佈圖
散佈圖加入類別



正關聯、負關聯

- **正關聯(Positive associated)**
  - 變數 X：高於(低於)平均值
  - 變數 Y： 高於(低於)平均值

- **負關聯(Negative associated)**
  - 變數 X： 高於(低於)平均值
  - 變數 Y： 低於(高於)平均值

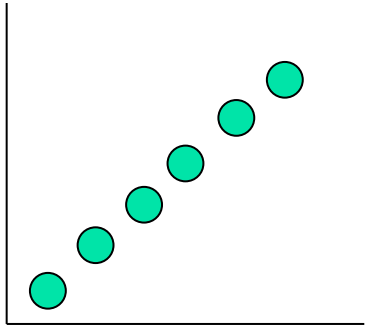
相關性(Correlation)

- **皮爾森相關係數(Pearson Correlation Coefficient)**

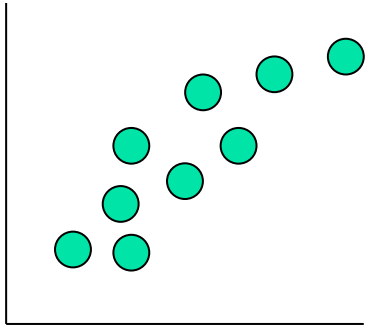
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

- **性質**
  - **皮爾森相關係數**的範圍介於 1~ -1 之間
  - $r > 0$：正相關
  - $r = 0$：無線性相關
  - $r < 0$：負相關

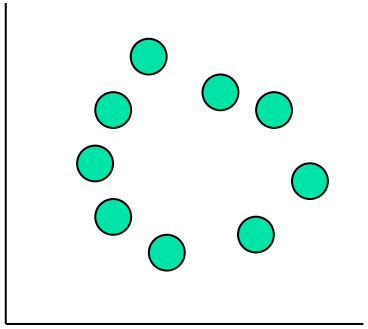
皮爾森相關係數與散布圖關係



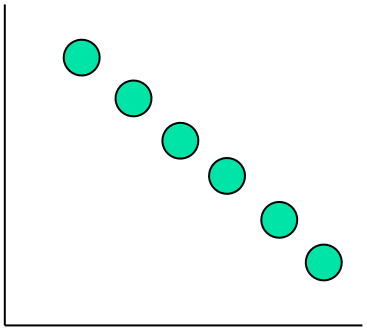
$r = +1$



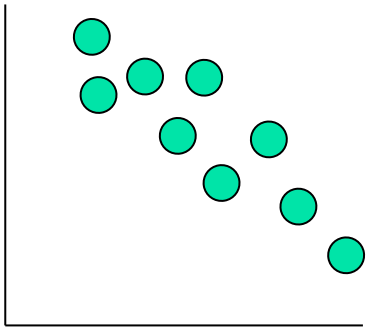
$0 < r < 1$



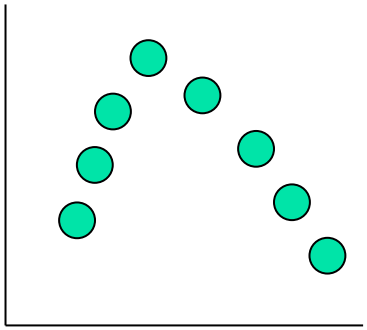
$r = 0$



$r = -1$



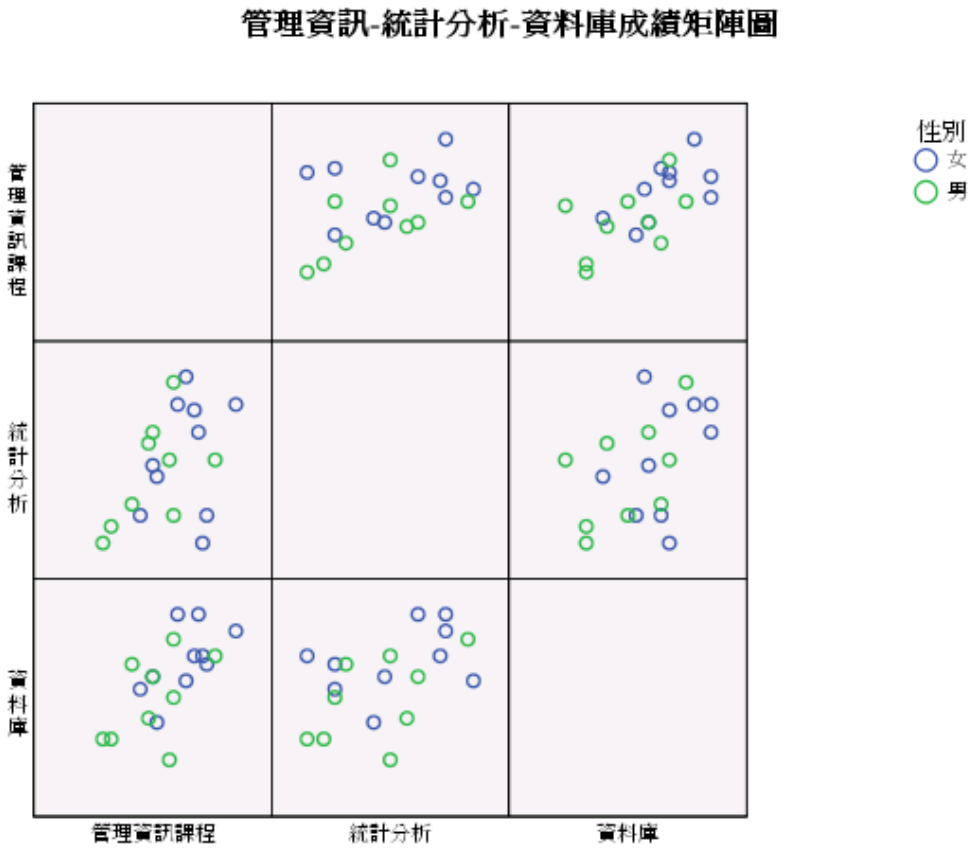
$-1 < r < 0$



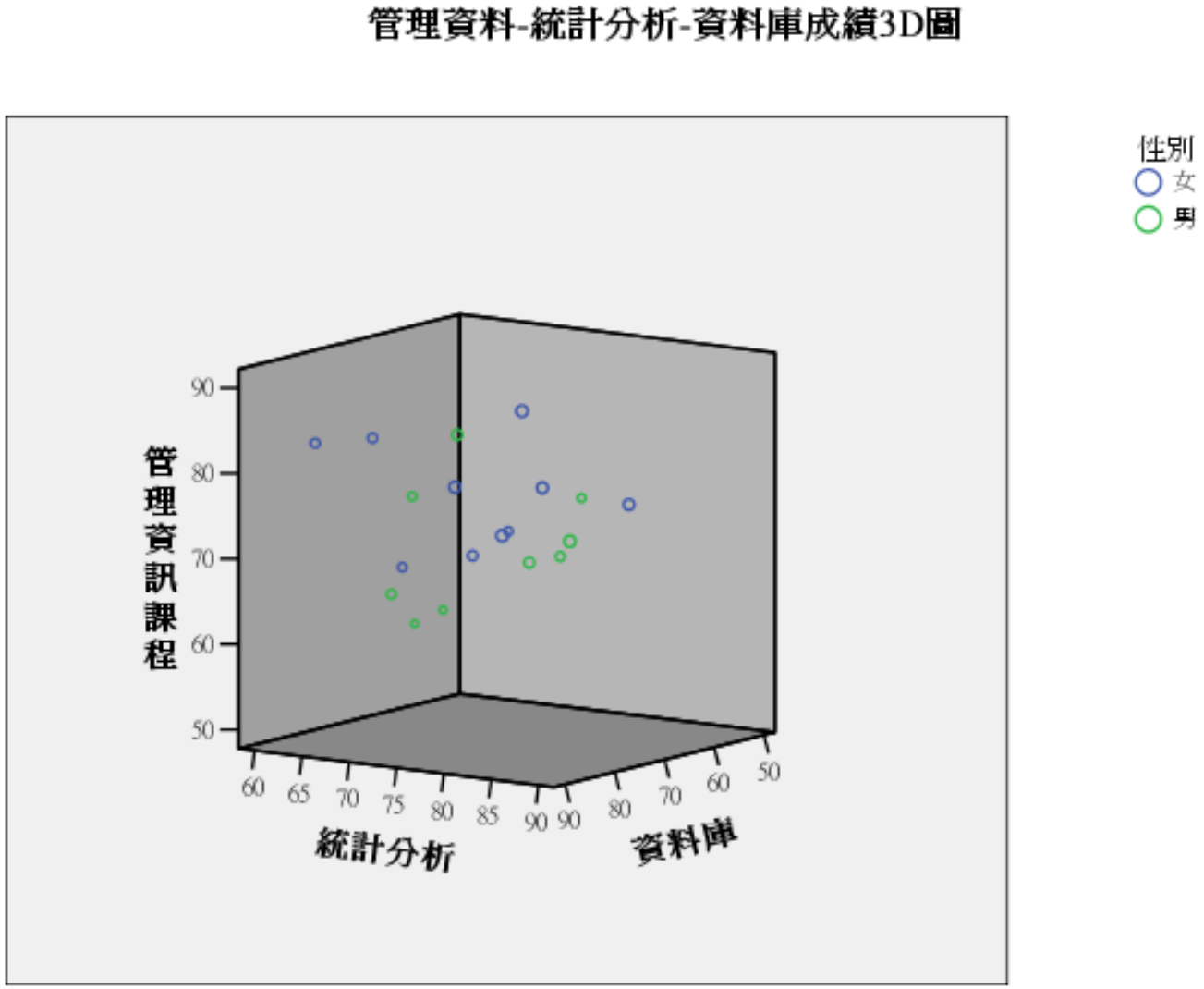
$r = 0$

散佈圖矩陣(Matrix Scatter Plot)

- **散佈圖矩陣**由John and Paul Turkey 提出，它能讓人一眼就看到所有兩兩變數的相關性。

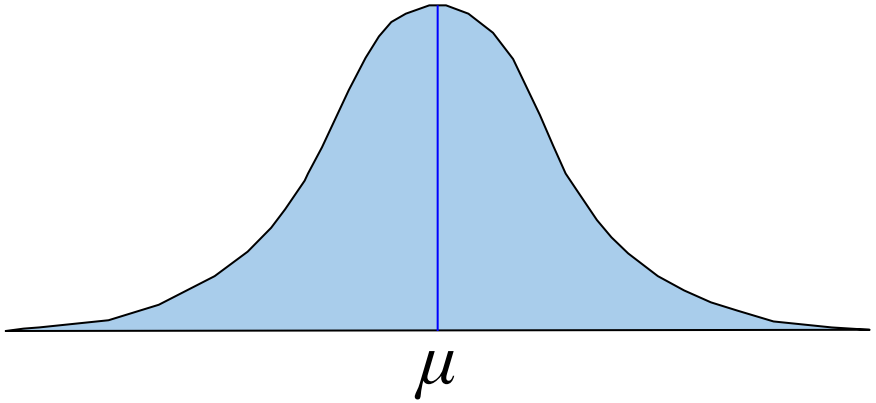


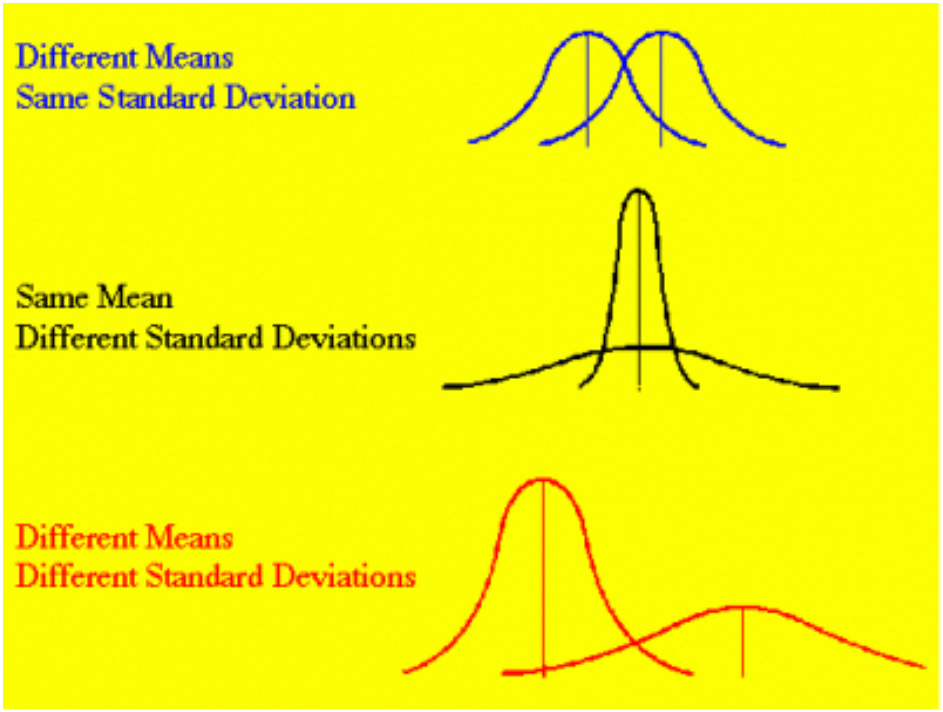
三維散佈圖(3D Scatter Plot)

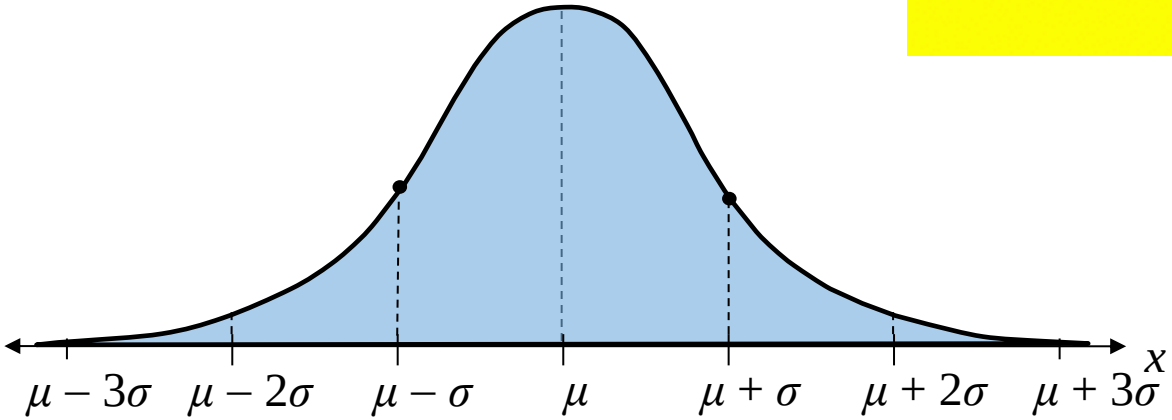


常態分布(Normal Distribution)

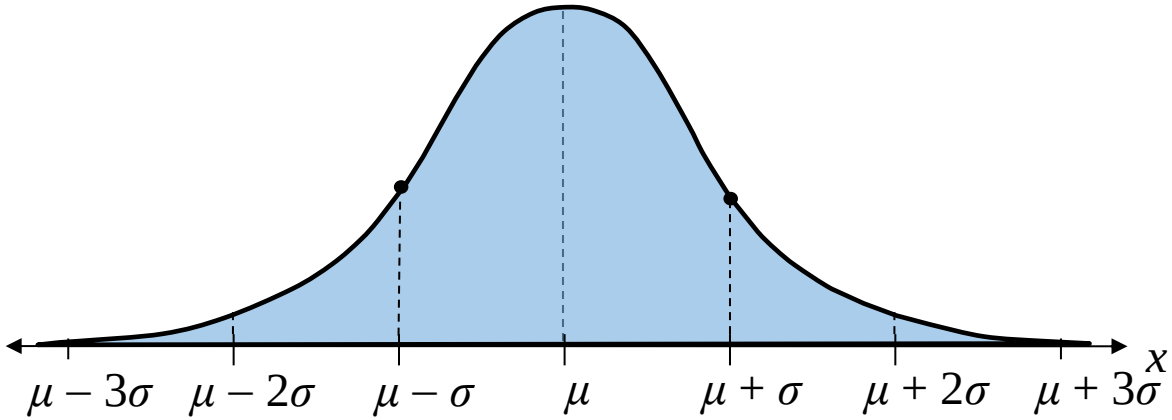
Normal Distribution





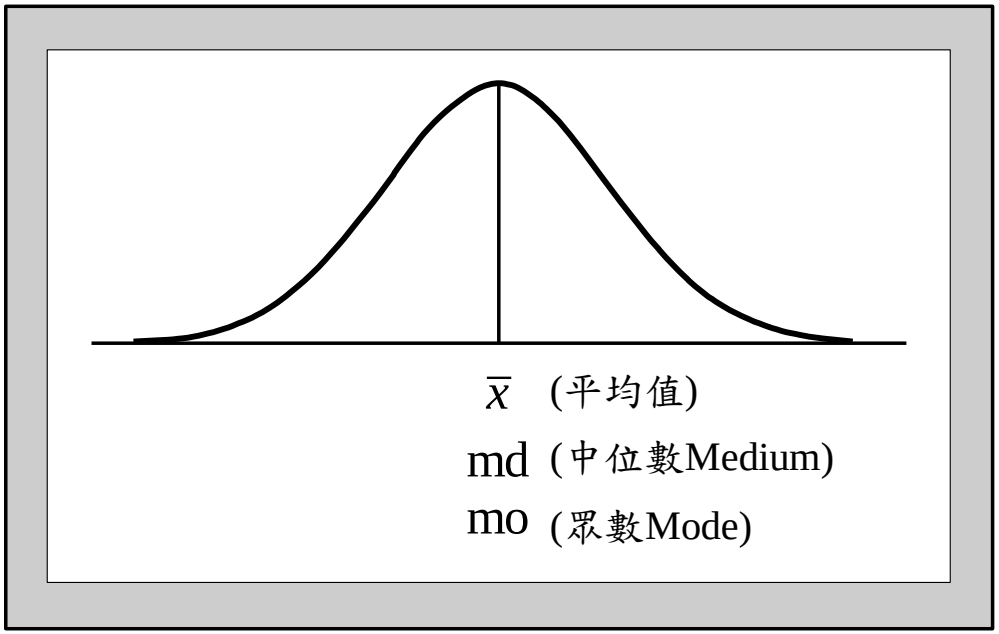


經驗法則(The Empirical Rule)
又稱 68%-95%-99.73%法則



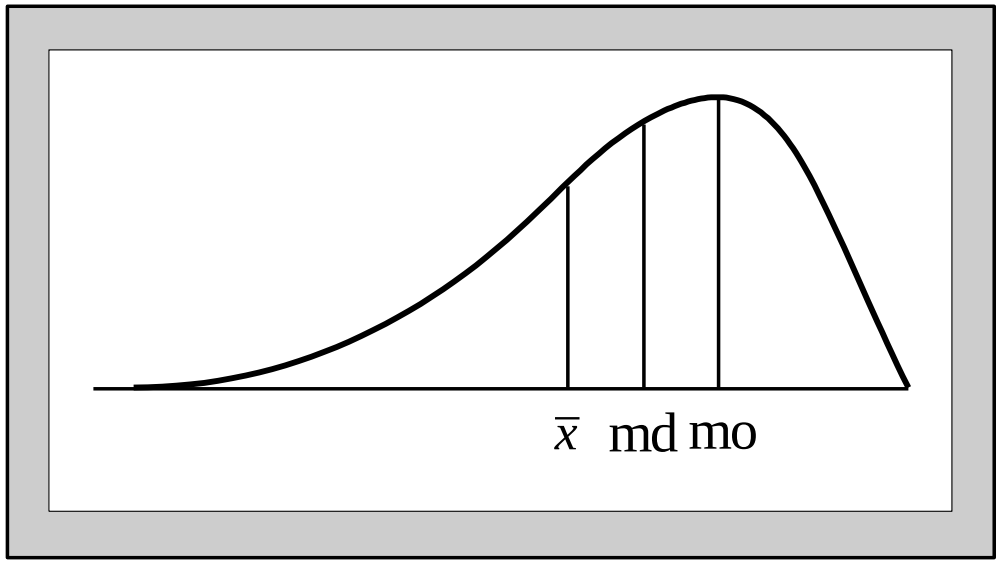
- **經驗法則限制條件**
  - 常態分配、單峰、對稱分布、鍾型分配
- **經驗法則**
  - **68%** 的資料落在 **μ±σ** 範圍內
  - **95%** 的資料落在 **μ±2σ** 範圍內
  - **97.7%** 的資料落在 **μ±3σ** 範圍內

單峰對稱分布 (Unimodal Symmetric Distribution)



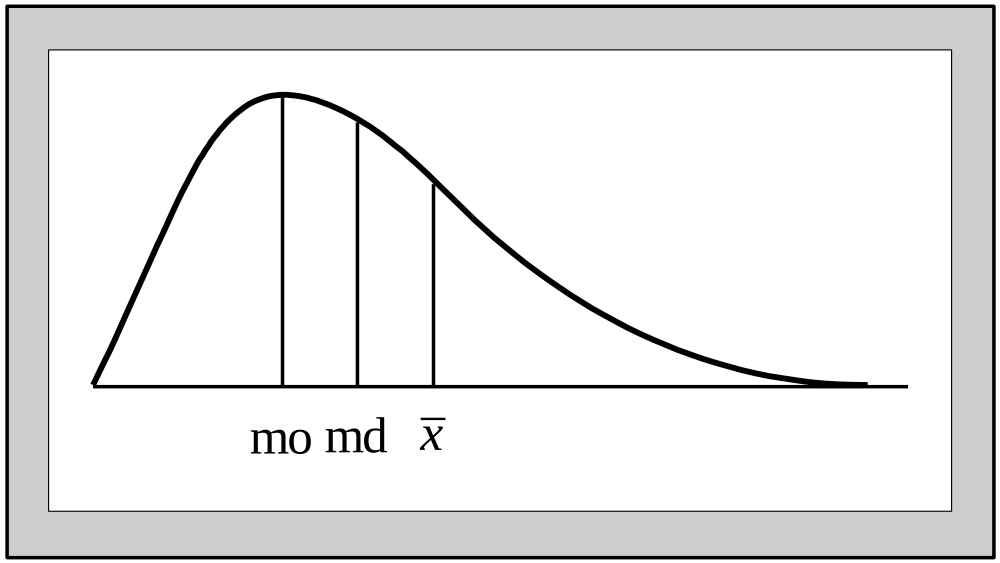
- **單峰對稱分布**
  - 平均值 = 中位數 = 眾數

單峰左偏分布(Unimodal Skewed to Left Distribution)



- 單峰左偏分布
  - 平均值 < 中位數 < 眾數

單峰右偏分布(Unimodal Skewed to Right Distribution)



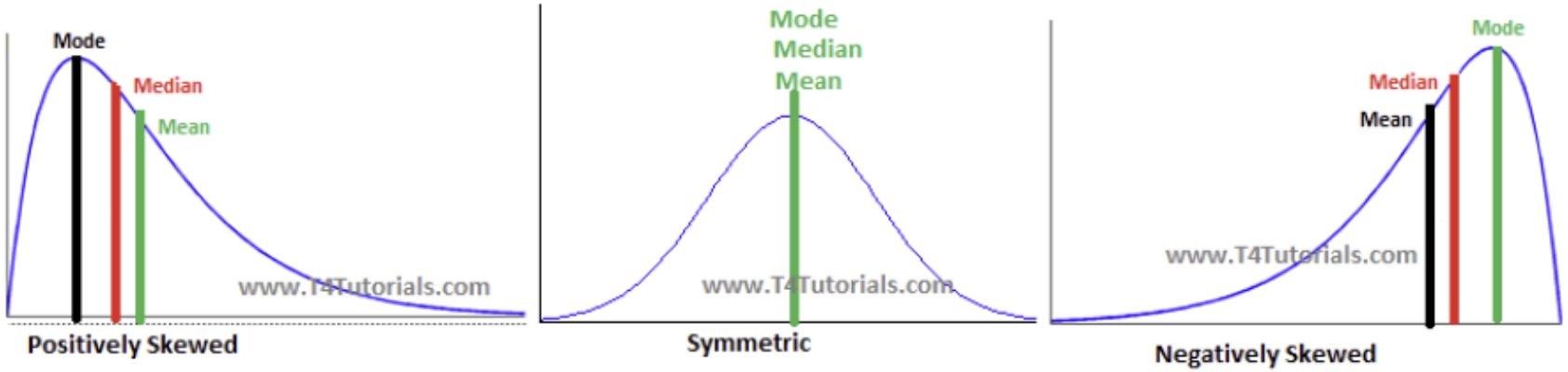
- 單峰右偏分布
  - 平均值 > 中位數 > 眾數

偏態係數(Skewness Coefficient)

- 偏態係數公式：

$$g_1 = \frac{m_3}{{m_2}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}}$$

$g_1 = 0$： 對稱分佈
$g_1 > 0$： 右偏分佈
$g_1 < 0$： 左偏分佈



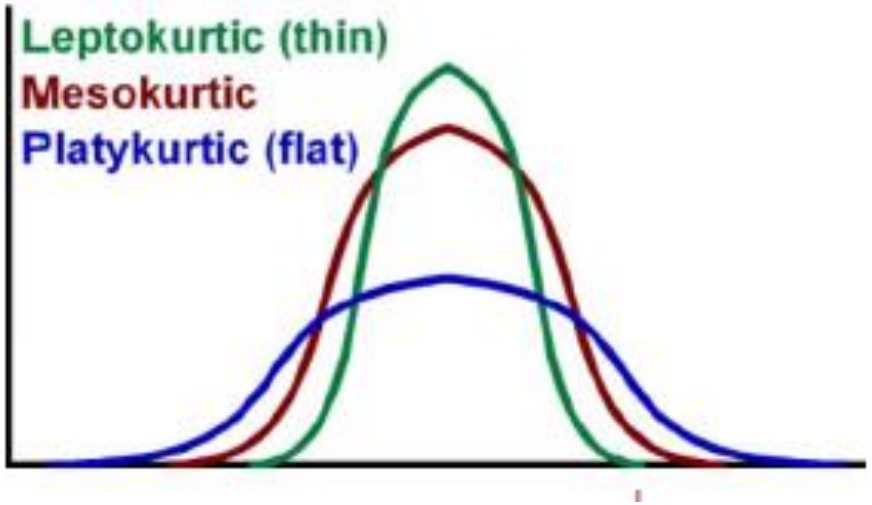
https://t4tutorials.com/data-skewness-in-data-mining/

峰態係數(Kurtosis Coefficient)

- **峰態係數**用來量測峰度高低的量數，可以反映資料的分佈形狀較為**高聳**或是**扁平**。

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3$$

$g_2 > 0$： 高峽峰(Lepto Kurtosis)
$g_2 = 0$： 常態峰(Normal Kurtosis)
$g_2 < 0$： 低潤峰 (Platy Kurtosis)



機率密度函數 (Probability Density Function；PDF)

- 給一個平均值μ以及標準差σ，我們可畫出一個常態分佈。
- 可以根據下列的**機率密度函式**公式，計算任意實數x_i發生在這個分佈的機率。

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

相對離差量數

- **相對離差量數**
  - 變異係數(Coefficient of Variation；CV)
  - 標準分數 Z-Score (Standard Score, Z-score)

- **變異係數**
  - CV = (標準差/平均值) × 100% $= (\sigma/\mu) \times 100\%$

標準分數Z-Score (Standard Score, Z-score)

- **標準分數(Standard Score)**，又稱 **Z-score**，中文稱為**Z-分數**或**標準化值**。
- 標準分數可藉由以下公式求出：

$$z = \frac{x - \mu}{\sigma}$$

- **Z -score 特性**
  - Z-score < 0：原始分數 *x* **小於**平均數
  - Z-score = 0：原始分數 *x* **等於**平均數
  - Z-score > 0：原始分數 *x* **大於**平均數

離差量數 (Measure of Dispersion)

- **絕對離差量數**
  - 全距(Range)
  - 四分位距(Interquartile-range；IQR)
  - 四分位差(Quartile Deviation；QD)
  - 平均差(Mean Deviation)
  - 標準差與變異數(Standard Deviation & Variance)

- **相對離差量數**
  - 變異係數(Coefficient of Variation；CV)
  - Z-Score

教師資訊

- 姓名：吳政瑋 (小吳老師)
- 現職：宜大資工專案助理教授
- 學歷：成功大學資工博士
- 研究興趣：資料探勘、人工智慧、AIoT應用
- 通訊方式
  - 電子信箱：wucw@niu.edu.tw
  - 校內電話: (03)9317331
  - Line: silvemoonfox
  - Office: 格致大樓E405室
  - 數位學習園區
- 實驗室：AI與資料科學實驗室
  - https://sites.google.com/view/cwwuadslab/

意見交流

歡迎提供意見與指導!!

您的寶貴意見將使本系更進步!!