第三章　數值敘述的技巧

- 中央趨勢之測量數
- 分數度之測量數
- 相對位置之測量數
- 形狀之衡量測量數
- 兩變數線性相關之測量數

3.1 中央趨勢之測量數

- 平均數(mean)
- 中位數(median)
- 眾數(mode)

3.1.1 平均數(mean)

本單元所討論的平均數係指算術平均數。

- 母體平均數

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

- 樣本平均數

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- 計算以分組資料的樣本平均數

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_k}$$

其中x_i及f_i分別為第 i 組的組中點及組次數。

● 加權平均數

$$\bar{X} = \frac{W_1 x_1 + W_2 x_2 + \cdots + W_n x_n}{W_1 + W_2 + \cdots + W_k}$$

其中x_i為觀察值，W_i為觀察值x_i的加權值

● 平均數的特性

給定一組樣本資料$x_1, x_2, \ldots, x_n$，則

1. $\sum_{i=1}^{n} x_i = n\bar{x}$
2. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
3. 對任意常數 T 滿足

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - T)^2$$

4. 優點：可用於統計推論，易進行代數運算，代表性強，且敏感度高。
5. 缺點：易受極端值影響。

3.1.2 中位數(median)

給定一組資料$x_1, x_2, \ldots, x_n$，求中位數 me

步驟 1：將上述資料由小至大排列成

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

步驟 2：

$$me=\begin{cases} X_{(\frac{n+1}{2})}，\text{n 是奇數} \\ \frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}]，\text{n 是偶數} \end{cases}$$

- 中位數的特性

1.給定一組資料$x_1, x_2, \ldots, x_n$，令中位數為 me，則任意常數 T 滿足

$$\sum_{i=1}^{n}|x_i-me| \leq \sum_{i=1}^{n}|x_i-T|$$

2.優點：不易受極端值影響，適合無母數統計推論。

3.缺點：只考慮居中一、二個資料，缺乏敏感度。

3.1.3 眾數(mode)

資料中出現次數最多之觀測值或類別，稱為眾數。

- 眾數的特性

1.優點：適用量的資料，亦適用質的資料。

2.缺點：一組資料的眾數是不一定，且敏感度較低。

3.2 分散度之測量數

3.2.1 全距(range)

R=最大值-最小值

- 全距的性質

1.優點：計算簡單，容易解釋。

2.缺點：易受極端值影響且準確度較低

3.2.2 四分位距(interquartile range)及四分位差(quartile devation)

$\mathrm{IQR} = Q_3 - Q_1$稱為四分位距

$QD = \frac{Q_3 - Q_1}{2}$稱為四分位差

Q_1及Q_3分別為第 1 及第 3 四分位數

3.2.3 平均絕對離差(mean absolute deviation)

給定一組資料$x_1, x_2, \ldots, x_n$，則

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- 平均絕對離差的特性
  1. 優點：觀念簡單，易於了解，代表性強。
  2. 缺點：涉及絕對值較不易計算，受極端值影響。

3.2.4 變異數(variance)及標準差(standard deviation)

母體變異數 $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$

母體標準差 $\sigma = \sqrt{\sigma^2}$

樣本變異數 $S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

樣本標準差 $S = \sqrt{S^2}$

- 變異數及標準差的特性

1.變異數及標準差的值恆大於等於零，其值越大表示資料越分散於平均數外，反之其值越小，表示資料越靠近於平均數。

2.優點：考慮到所有觀察值的資訊，適合代數運算。

3.缺點：受極端值影響。

3.2.5 變異係數(coefficient of variation 以 CV 表示)

母體變異係數 $\mathrm{CV} = \frac{\sigma}{\mu} * 100\%$

樣本變異係數 $\mathrm{cv} = \frac{s}{x} * 100\%$

- 變異係數的特性

  兩組或兩組以上資料，當單位不同或單位相同但平均數差異大時，可用變異細數比較分散程度。

3.3 位置之衡量統計量

3.3.1 百分位數(percentiles)

給定一組樣本資料$x_1, x_2, \ldots, x_n$，求第 k 個百分數P_k

步驟 1：先將樣本資料由小至大排列成$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

步驟 2：求位置$r = \frac{k}{100} * n$

步驟 3：

$$P_k = \begin{cases} \frac{1}{2}\{X_{(r)} + X_{(r+1)}\}, & \text{r 為整數} \\ X_{([r+1])}, & \text{r 不為整數} \end{cases}$$

[ ]為高斯符號

3.3.2 四分位數(quartiles)

$q_1 = P_{25}$

$q_2 = P_{50}$

$q_3 = P_{75}$

3.3.3 Z 值

定義：$Z_i = \frac{x_i - \mu}{\sigma}$ 或 $Z_i = \frac{x_i - \bar{x}}{s}$

其中Z_i表示第 i 項觀測值的 Z 值

$\bar{x}$或μ分別代表樣本或母體的平均數

s 或σ分別代表樣本或母體的標準差

3.4 形狀之衡量統計量

3.4.1 偏態係數

母體：$S_k = \frac{3(\mu - M_e)}{\sigma}$

樣本：$S_k = \frac{3(\bar{x} - m_e)}{s}$

M_e及m_e分別表示母體及樣本的中位數

1. $S_k = 0$表示分配成對稱型態

2. $S_k > 0$表示分配成右偏

3. $S_k < 0$表示分配成左偏

3.4.2 峰度係數

母體：$CK = \dfrac{\sum_{i=1}^{n} (x_i - \mu)^4 / N}{\sigma^4}$

樣本：$ck = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{s^4}$

1.CK=3 分配呈常態峰

2.CK>3 分配呈高峽峰

3.CK<3 分配呈低闊峰

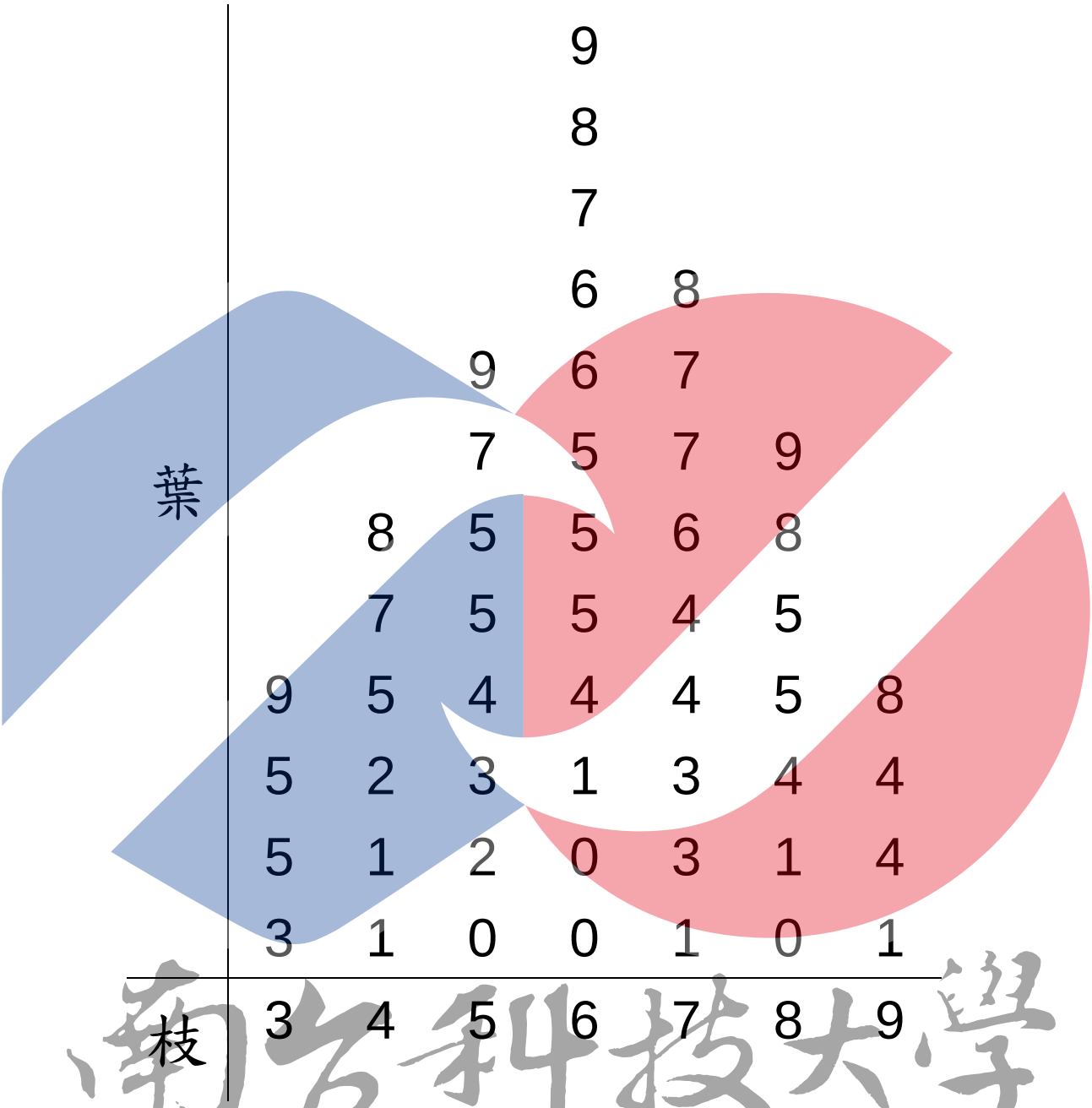
3.6 探索性資料分析(exploratory data analysis：EDA)

- 枝葉圖(stem-and-leaf diagrams)
- 箱型圖(box and whisker plots)

3.6.1 枝葉圖

以例 2.3 50 名學生統計學成績說明枝葉圖

枝	葉									
3	3	5	5	9						
4	7	1	5	1	8	2				
5	2	7	0	4	5	9	3	5		
6	0	5	5	4	8	1	9	7	0	6	6	5
7	7	4	1	4	3	7	8	6	3	
8	5	4	1	0	5	9	8			
9	1	4	8	4						



順序枝葉圖

枝葉圖類似長條圖又保留資料的數值，提供更多資訊。

3.6.2 盒鬚圖

給定一組資料$x_1, x_2, \ldots, x_n$，試繪出盒鬚圖，並判定資料呈對稱、右偏或左偏，同時辨認觀察值是否為異常值?

步驟 1: 求出樣本資料之最小值,q_1, q_2, q_3及最大值

步驟 2：

求出左邊兩條圍籬

$$l_1: q_1 - 3(q_3 - q_1)$$

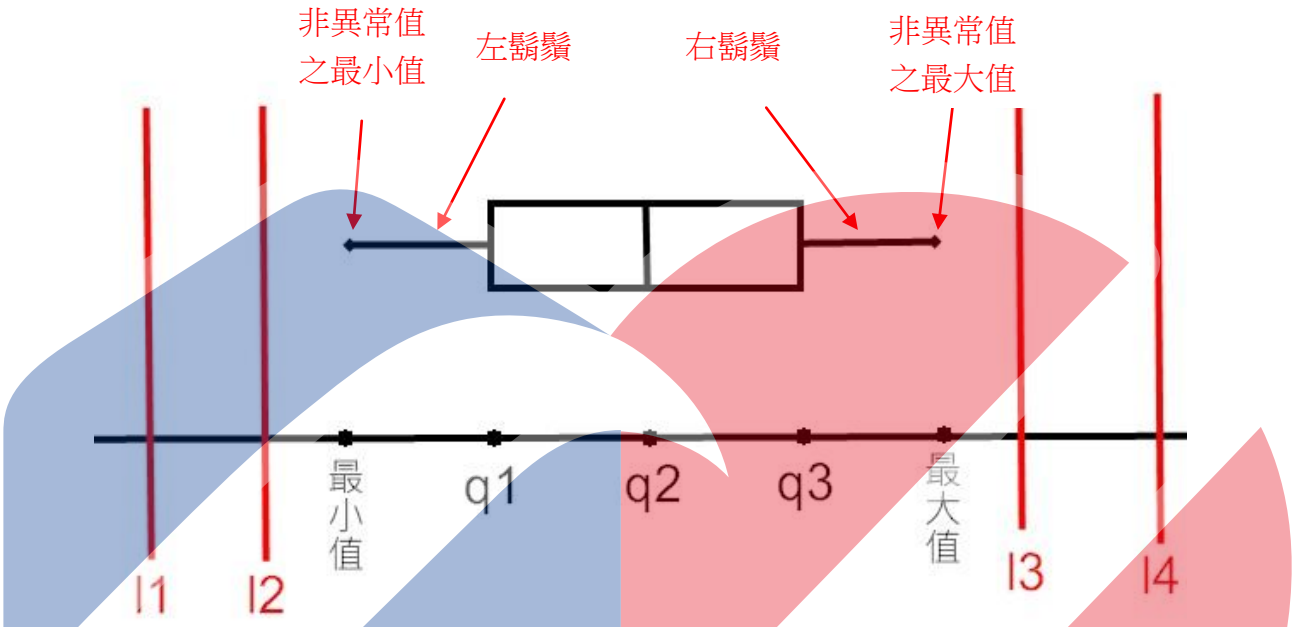
$$l_2: q_1 - 1.5(q_3 - q_1)$$

求出右邊兩條圍籬

$$l_3: q_3 + 1.5(q_3 - q_1)$$

$$l_4: q_3 + 3(q_3 - q_1)$$

步驟３：畫出盒鬚圖



步驟４：

左鬍鬚>右鬍鬚　資料呈左偏

左鬍鬚<右鬍鬚　資料呈右偏

左鬍鬚=右鬍鬚　資料呈對稱

步驟 5：觀察值落在l_1與l_2之間或l_3與l_4之間稱為可疑異常值，觀察直落在l_1左邊或l_4右邊稱為異常值。

3.7 兩變數線性相關之測量數

- 共變異數(covariance)
- 相關係數(coefficient of correlation)

母體共變異數

$$\mathrm{COV}(X,Y)=\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$$

μ_x及μ_y分別代表隨機變數 X 及 Y 之母體平均數

樣本共變異數

$$\mathrm{cov}(x,y)=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$$

$\bar{x}$及$\bar{y}$分別代表隨機變數 x 及 y 之樣本平均數

1.cov(x,y)>0 表示變數 X 與 Y 為正相關

2.cov(x,y)<0 表示變數 X 與 Y 為負相關

3.cov(x,y)=0 表示變數 X 與 Y 無線性相關

相關係數

母體相關係數 $\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$

COV(X,Y)表示母體共變異數

σ_x及σ_y分別表示隨機變數X及Y之母體標準差

樣本相關係數 $r = \frac{cov(x,y)}{S_x S_y}$

cov(x,y)表示樣本共變異數

S_x及S_y分別表示隨機變數x及y之樣本標準差

1.cov(x,y)>0，則 r>0 表示變數x與y正線性相關

2.cov(x,y)<0，則 r<0 表示變數x與y負線性相關

3.cov(x,y)=0，則 r=0 表示變數x與y無線性相關

cov(x,y)的值在(-∞,∞)之間，無法從數值大小判斷相關程度，r值在[-1,1]之間，由相關係數r的大小及正負可知變數X與Y相關程度的強度與方向。