第三章 敘述統計(II) 統計量數

本章內容

3-1 集中趨勢量數

3-2 位置參數

3-3 分散趨勢量數

3-4 資料的偏態

3-5 謝比雪夫不等式與經驗法則

3-6 探索式資料分析

前一章介紹統計圖表的製作，好的統計圖表方便閱讀者在很短的時間內獲得關於資料中所喻含的中要訊息。我們觀察這些統計分配圖形，概略可從三種特徵去表現圖形的樣態，換言之，倘若我們知道這三個特徵值，也了解每一個特徵值在圖形上的意義，就可以概略獲悉資料分配的樣態。這三種特徵為

1. 集中趨勢 (central tendency)：描述大部分資料座落的位置，呈現資料分配之中心位置或共同趨勢。例如：大多數新生兒的體重與身高的同學成績
2. 分散趨勢(dispersion tendency)：描述一組資料內，彼此間的差異程度。
3. 形狀參數(shape parameters)：一組資料的分配形狀有很多樣態，大部分的情形資料都是單峰的，也就是平滑後的分配只有一個山峰的形狀，就一個山峰的形狀，經驗上可以用兩種形狀參數來表現出它們間的的差異。
    - **偏態(skewness)**：所謂山峰乃中間高，兩邊漸低，分配曲線往兩邊下滑的幅度差異造成形狀上的差異。若像右下滑較為緩慢，我們稱為右偏。若向左下滑較為緩慢，則稱為左偏。左右近似相同，稱為對稱分配。
    - **峰態(kurtosis)**：山峰除了偏向外，亦有陡峭之別。陡峭的分配稱為高狹峰，平坦的山峰稱為平闊峰。

3-1 集中趨勢量數

集中趨勢量數的目地為呈現資料分配之中心位置或共同趨勢，本節介紹三種常用的統計量數。

3.1.1 平均數

平均數(mean, average)又稱算術平均數。假設有一組母體資料為 $x_1,\cdots,x_N$ ，母體平均值以 μ 表之，定義為

$$\mu=\frac{x_1+\cdots+x_N}{N}=\frac{\sum_{i=1}^{N}x_i}{N} \qquad \text{式 3-1}$$

例 3-1

2008 年， 20 位新入學皮卡丘幼稚園小班幼兒的體重為

18 17 18 16 17 15 15 16 18 15

14 16 12 14 13 16 15 19 16 16

$\mu=(18+17+\cdots+16)/20=15.8$

R 程式語法：

```
#首先輸入資料, 放在程式變數 x
> x<-c(18, 17, 18, 16, 17, 15, 15, 16, 18, 15, 14, 16, 12, 14, 13, 16, 15, 19, 16, 16)
> mean(x)  #mean 是 R 的一個函式, 傳回資料的算術平均數, 通常函式以原意的英文或其縮寫表示。
[1] 15.8
```

例 3-2

利用程式計算表 2-1 晶圓片的平均厚度。

$\mu=(1516+\cdots+1461)/50=1491.02$

從一個母體中隨機抽取 n 個樣本 $x_1,\cdots,x_n$ ，樣本平均值以 $\bar{x}$ 表之，定義為

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$ 式 3- 2

例 3- 3

隨機抽出 5 位皮卡丘幼稚園小班兒童，測量其體重分別為 15, 18, 16, 19, 18

$$\bar{x} = \frac{15+18+16+19+18}{5} = 17.2$$

再隨機抽出 5 位，測量其體重分別為 15, 14, 15, 18, 12

$$\bar{x} = \frac{15+14+15+18+12}{5} = 14.8$$

樣本平均值會隨著每次抽樣所選取的樣本不同而改變。

平均數的性質：(代數性質)

1. 假設 A 母體的個數為 N_A，其平均值為 μ_A；B 母體的個數為 N_B，其平均值為 μ_B。將兩母體合併，母體個數為 $N = N_A + N_B$，母體平均數為

$$\mu = \frac{N_A\mu_A + N_B\mu_B}{N_A + N_B}$$ 式 3- 3

2. 假設有 k 個母體，母體個數與平均值分別為 $N_1, \cdots, N_k$ 和 $\mu_1, \cdots, \mu_k$。將此 k 個母體合併母體個數為 $N = N_1 + \cdots + N_k = \sum_{i=1}^{k} N_i$，母體平均數為

$$\mu = \frac{\sum_{i=1}^{k} N_i\mu_i}{\sum_{i=1}^{k} N_i}$$

例 3- 4

假設 A 班 50 位同學品管平均成績為 78，B 班 60 位同學品管平均成績為 80，則兩班平均成績為

$$\mu_{A+B} = \frac{50\times 78 + 60\times 80}{50+60} = 79.09$$

3. 離差總和為零：離差(deviation)為觀察值與平均值之差，第 i 個觀察值得離差

為 $x_i - \bar{x}$，離差大於零表示此觀察值大於平均值，若小於零表示觀察值小於平均值。所有觀察值得離差總和為零。即

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

此表示平均值位於所有觀察值的重心位置，大於零(右側)與小於零(左側)的數恰好取得平衡，故離差總和為零。

例 3- 5

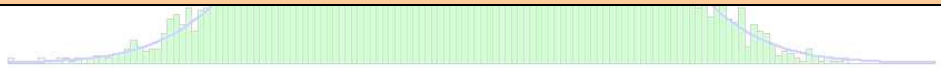
接續上

例 3- 3，觀察值的離差為

$$x_1 - \bar{x} = 15 - 17.2 = -2.2$$
$$x_2 - \bar{x} = 18 - 17.2 = 0.8$$
$$x_3 - \bar{x} = 16 - 17.2 = -1.2$$
$$x_4 - \bar{x} = 19 - 17.2 = 1.8$$
$$x_5 - \bar{x} = 18 - 17.2 = 0.8$$

右側離差和為 0. 8+1. 8+0. 8=3. 4

左側離差和為-2. 2+(-1. 2)=-3. 4, 故總和為零。



4. 離差平方和最小：**令** a 為任意一點，第 i 個觀察值與它的差為 $x_i - a$。它的平方總和大於或等於離差平方和；即

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{式 3- 4}$$

上式中等式成立的唯一條件為 $\bar{x} = a$。

例 3- 6

接續例 3- 1，令 $a = 17$，則對 a 的離差平方和為

$$\sum_{i=1}^{5}(x_i - 17)^2 = (-2)^2 + (1)^2 + (-1)^2 + (2)^2 + (1)^2 = 11$$

對平均數的離差平方和為

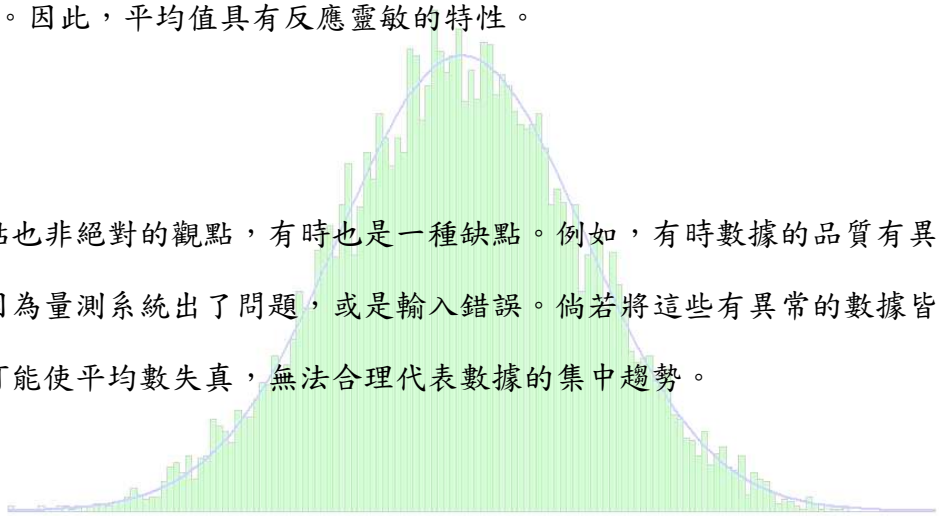
$$\sum_{i=1}^{5}\left(x_i-\bar{x}\right)^2=(-2.2)^2+(0.8)^2+(-1.2)^2+(1.8)^2+(0.8)^2=10.8$$

平均數的優點：

1. 所有觀察值皆列入計算。
2. 每個觀察值代表性，重要性相同。在計算公式上，每個觀察值的權數(weight)都相同為$1/n$，並沒有偏袒或忽視任何一個被觀察到數值，在計算算術平均數上沒有主觀，或其他考量，所以平均值的運用極為普遍。
3. 因為所有觀察值皆列入計算，所以觀察值上的任何改變都會反應平均值的計算結果。因此，平均值具有反應靈敏的特性。

註：

然而優點也非絕對的觀點，有時也是一種缺點。例如，有時數據的品質有異常，可能是因為量測系統出了問題，或是輸入錯誤。倘若將這些有異常的數據皆列入計算，可能使平均數失真，無法合理代表數據的集中趨勢。



3.1.2 中位數(median)

顧名思義此數在數據的位置排序上處於中間的位置。轉換成比較明確的說法，即有**至少一半的數據大於或等於此數，也有至少一半的數據小於或等於此數**。(可能有人會簡化說成 “一半的數據大於此數，也有一半的數據小於此數 ”，但這是錯誤的說法。)根據上述中位數之定義，當數據的個數為奇數時，中位數具有唯一性（僅有一個且唯一一個符合上述的定義)；然而，當數據的個數是偶數時，就可能會有無限多個數滿足上述定義。為了避免爭端，我們以計算方式去定義中位數。

觀察值$x_1,\cdots,x_n$的中位數為

$$Me=\begin{cases} x_{(\frac{n+1}{2})} & ，當n為奇數 \\ \dfrac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2} & ，當n為偶數 \end{cases}$$ 式 3- 5

讓 $x_{(1)},\cdots,x_{(n)}$ 是觀察值 $x_1,\cdots,x_n$ 從小排至大的有序統計量，其定義為

$$\begin{aligned} x_{(1)} &= \min\{x_1,\cdots,x_n\} \\ x_{(2)} &= \min\{x_1,\cdots,x_n\}\setminus\{x_{(1)}\} \\ &\vdots \\ x_{(n-1)} &= \min\{x_1,\cdots,x_n\}\setminus\{x_{(1)},\cdots,x_{(n-2)}\} \\ x_{(n)} &= \min\{x_1,\cdots,x_n\}\setminus\{x_{(1)},\cdots,x_{(n-1)}\}=\max\{x_1,\cdots,x_n\} \end{aligned}$$ 式 3- 6

函數 min{}為求最小值（minimum），max{}為求最大值。 集合運算元 \ 為扣除相同元素的動作。例如 $\{1,2,3,4\}\setminus\{3\}=\{1,2,4\}$ 。

例 3- 7

求 8, 9, 11, 12, 13 的中位數。

因為 $n=5$ 為奇數，所以中位數為 $x_{\left(\frac{5+1}{2}\right)}=x_{(3)}=11$

例 3- 8

求 1, 2, 3, 4 的中位數。

因為 $n=4$ 為偶數，所以中位數為 $\dfrac{x_{(2)}+x_{(3)}}{2}=\dfrac{2+3}{2}=2.5$

例 3- 9

五位男同學的期中考成績 64, 52, 70, 74, 68，六位女同學的期中考成績 55, 48, 90, 83, 77, 89。

男同學成績的中位數為 68，女同學成績的中位數為 $\dfrac{77+83}{2}=80$；男女同學成績的中位數為 70。

注意：$70 \neq \frac{68+80}{2}$。我們無法由兩組資料的中位數，利用代數運算的方法求得合併後資料的中位數。

中位數的性質：

1. 中位數離差絕對值和最小：各觀察值與任一值 a 之差的絕對值（$|x_i - a|$）總和大於或等於中位數離差值總和，即

$$\sum_{i=1}^{n} |x_i - a| \geq \sum_{i=1}^{n} |x_i - Me| \qquad \text{式 3- 7}$$

例 3- 10

接續前例，五位男同學的期中考成績 64，52，70，74，68，$Me = 68$。中位數離差絕對值為 4，16，2，6，0，總和 $\sum_{i=1}^{5} |x_i - Me| = 28$。

(a) 令 $a = 70$，則 $|x_i - a|$ 分別為 6，18，0，4，2，$\sum_{i=1}^{5} |x_i - a| = 30$。

(b) 令 $a = 72$，則 $|x_i - a|$ 分別為 8，20，2，2，4，$\sum_{i=1}^{5} |x_i - a| = 36$。

(c) 令 $a = 67$，則 $|x_i - a|$ 分別為 3，15，3，7，1，$\sum_{i=1}^{5} |x_i - a| = 29$。

例 3- 11

六位女同學的期中考成績 55，48，90，83，77，89，$Me = 80$。中位數離差絕對值為 25，32，10，3，3，9，總和 $\sum_{i=1}^{6} |x_i - Me| = 82$。

(a) 令 $a = 81$，則 $|x_i - a|$ 分別為 26，33，9，2，4，8，$\sum_{i=1}^{6} |x_i - a| = 82$。

(b) 令 $a = 90$，則 $|x_i - a|$ 分別為 35，42，0，7，13，1，$\sum_{i=1}^{6} |x_i - a| = 98$。

2. 不易受極端值影響。

例 3- 12

計算下列兩組數據的中位數與平均數。

(a) 14, 13, 17, 18, 20

(b) 14, 13, 17, 18, **200**

解：(a) $Me=17, \bar{x}=16.4$ (b) $Me=17, \bar{x}=52.4$

3.1.3 眾數(mode)

眾乃多之意，即出現次數最多者稱之。所以，一組數據中某一個數字出現的次數超過一次，且是所有出現數字中重覆最多次者稱為眾數。

例 3- 13

擲骰子 12 次，結果為 4, 2, 1, 6, 5, 3, 3, 2, 4, 3, 6, 3。次數分配表如下：

點數	1	2	3	4	5	6
次數	1	2	4	2	1	2

其中點數 3 出現最多次，所以眾數為 3。

例 3- 14

求以下三組數據的眾數。

(A) 15, 17, 15, 12, 14,

(B) 14, 14, 2, 4, 2, 3, 1, 9

(C) 1, 2, 3, 4, 5

解答：

(A)15 (B) 14 和 2 (C) 無

註：眾數可能不只一個，或不存在。

3-2 位置參數

3.2.1 百分位數

前面我們的定義中位數為“**一半以上的數據大於或等於此數，也有一半以上的數據小於或等於此數**”，概念上中位數將一組資料分成兩等份。我們將這樣的概念對廣，我們取 99 個分割點可以將資料分割成 100 等份，我們稱這些分割點為百分位數(percentile)。這百分位數可以讓我們知道一個樣本點(sample point)在全體數據的相對排名順序。

定義：

令 P_i 代表一組樣本的第 i 個百分位數，在這組樣本中“至少有 $(100-i)\%$ 的數據大於或等於此數，也至少有 $i\%$ 的數據小於或等於此數”

第 i 個百分位數 P_i 的計算方法。

1. **將資料排序得到有序統計量 $x_{(1)},\cdots,x_{(n)}$ 。**
2. 求出位置指標 k 。

$$k=\frac{i}{100}\times n$$

3.

(i) **若 k 為整數，** $P_i=\dfrac{x_{(k)}+x_{(k+1)}}{2}$

(ii) **若 k 為非整數，** $P_i=x_{([k])}$ **，其中 $[k]$ 代表大於 k 的最小整數。例如：**

[4.3]=5。

3.2.2 十分位數

顧名思義，十分位數(decile)乃是利用 9 個分割點將資料分割成 10 等分。這 9 個分割點為 $D_1, D_2, \cdots, D_9$。事實上，根據百分位數的定義，我們可以得出

$$D_1 = P_{10}, D_2 = P_{20}, \cdots, D_9 = P_{90}。$$

3.2.3 四分位數

同樣的，四分位數(quartile)乃是利用 3 個分割點將資料分割成 4 等分。這 3 個分割點為 Q_1, Q_2, Q_3。事實上，根據百分位數的定義，我們可以得出

$$Q_1 = P_{25}, Q_2 = P_{50}, Q_3 = P_{75}。$$

所以，同學只要熟悉百分位數的定義與計算，就可以得出十分位數，四分位數和中位數。

例 3- 15

計算表 2-1，晶圓厚度 3 個四分位數與 P_{30}。

首先，將資料排序如下，

1379 1385 1412 1425 1427 1428 1448 1449 1451 1455

1459 1460 1460 1461 1461 1462 1465 1469 1470 1470

1472 1474 1482 1483 1483 1484 1488 1495 1500 1500

1501 1505 1509 1512 1512 1513 1516 1517 1520 1533

1543 1546 1547 1548 1554 1566 1568 1569 1601 1614

(1) 第一個四分位數 Q_1。

因為 $Q_1 = P_{25}$，$i = 25$，位置指標 $k = \frac{25}{100} \times 50 = 12.5$，所以，$Q_1 = x_{(13)} = 1460$。

(2) 第二個四分位數 Q_2。

因為 $Q_2 = P_{50}$，$i = 50$，位置指標 $k = \frac{50}{100} \times 50 = 25$，所以，

$$Q_2=\frac{x_{(25)}+x_{(26)}}{2}=\frac{1483+1484}{2}=1483.5\ 。$$

(3) 第一個四分位數 Q_3。

因為 $Q_3=P_{75}$，$i=75$，位置指標 $k=\frac{75}{100}\times 50=37.5$，所以，$Q_3=x_{(38)}=1517$。

(4) P_{30}。

$i=30$，位置指標 $k=\frac{30}{100}\times 50=15$，所以，$P_{30}=\frac{x_{(15)}+x_{(16)}}{2}=\frac{1461+1462}{2}=1461.5$

3-3 分散趨勢量數

分散趨勢量數，顧名思義，乃在衡量一組資料內部彼此間的差異程度。通常以兩個數間的距離代表兩個數的差異程度。而一組資料，描述這組資料內容的差異程度，一種作法是合計每一個數與集中趨勢量數(中位數或平均值)的距離。另一作法則是以位置參數間的距離代表全體資料的差異程度。

採用第一種作法的有變異數(variance)，標準差(standard deviation)和平均絕對離差(mean absolute deviation)。採用第二種作法的有全距(range)，四分位距，四分位差。

3.3.1 變異數

變異數是合計每一數值與平均值間的差異程度。其定義如下:

1. 母體變異數：假設有一組母體資料為 $x_1,\cdots,x_N$，母體平均值為 μ 表之，母體變異數為

$$\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \qquad \text{式 3- 8}$$

這裡我們用 $(x_i-\mu)^2$ 代表第 i 個數據與母體平均值的距離。所以，母體變異數 σ^2（讀成/sigma/平方）為這些距離的平均值。

2. 樣本變異數：假設有一組樣本資料為 $x_1,\cdots,x_n$，樣本平均值為 $\bar{x}$，樣本變異數

定義為

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} \qquad \text{式 3-9}$$

這裡我們注意到樣本變異數的計算與母體變異數略有不同，上式的分母為樣本數減一，其原理屬數理統計範疇，再次略去不討論，同學僅需注意公式上不同，在計算上不要犯錯。

例 3- 16

接續例 3- 1，計算 20 位皮卡丘幼稚園小班幼兒體重的母體變異數

18 17 18 16 17 15 15 16 18 15

14 16 12 14 13 16 15 19 16 16

解：

1. 計算母體平均值

$$\mu = (18+17+\cdots+16)/20 = 15.8$$

2. 計算離差

2.2 1.2 2.2 0.2 1.2 -0.8 -0.8 0.2 2.2 -0.8

-1.8 0.2 -3.8 -1.8 -2.8 0.2 -0.8 3.2 0.2 0.2

3. 計算變異數

$$\sigma^2 = \frac{2.2^2+\cdots+0.2^2}{20} = 2.96$$

例 3- 17

接續例 3- 9，計算五位男同學期中考成績的樣本變異數。

解：

1. 計算樣本平均值

$$\bar{x} = \frac{64+\cdots+68}{5} = 65.6$$

2. 計算離差　　-1.6 -13.6　4.4　8.4　2.4

3. 計算變異數

$$s^2=\frac{(-1.6)^2+\cdots+2.4^2}{4}=70.8$$

練習：接續例 3- 9，計算六位女同學期中考成績的變異數。

標準差：

因為變異數使用離差平方，所以變異數的單位為資料單位的平方。例如，身高的單位為公分，則身高變異數的單位為平方公分。為了使統計量數的單位與觀察值的單位相同，所以，提出將變異數開根號的方式來衡量資料的分散程度，稱為標準差。

母體標準差

$$\sigma=\sqrt{\sigma^2}=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$ 式 3-10

樣本標準差

$$s=\sqrt{s^2}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$ 式 3-11

例 3- 18

計算 20 位皮卡丘幼稚園小班幼兒體重的母體標準差。

$$\sigma=\sqrt{2.96}=1.72$$

計算 5 位男同學體重的樣本標準差

$$s=\sqrt{70.8}=8.41$$

例 3- 19

利用程式計算表格 2-1 晶圓厚度的變異數與標準差。

```
> xx
 [1] 1516 1512 1472 1449 1533 1505 1501 1500 1455 1554 1448 1462 1482
1412 1500
[16] 1568 1470 1509 1460 1483 1428 1427 1425 1517 1460 1512 1543 1488
1513 1385
[31] 1601 1470 1459 1451 1569 1484 1548 1520 1474 1566 1483 1547 1379
1461 1469
[46] 1495 1465 1614 1546 1461
> var(xx) # 計算變異數的函數 variance
[1] 2541.163
> sd(xx) #計算標準差的函數 standard deviation
[1] 50.40995
```

3.3.2 平均絕對離差

數學上計算距離的方式，除了離差平方外，絕對離差 (absolute deviation)$\left|x_i - \bar{x}\right|$ 也是常用的方式之一。平均絕對離差 (mean absolute deviation, *MAD*)就是將各觀察值之絕對離差的平均值，定義為

$$MAD = \frac{\sum_{i=1}^{n} \left|x_i - \bar{x}\right|}{n} \qquad \text{式 3-12}$$

若是計算母體的平均絕對離差，將上式樣本平均值 $\bar{x}$ 用母體平均值取代之。

例 3-20

接續例 3-9，計算五位男同學期中考成績的平均絕對離差。

解：

1. 計算樣本平均值

$$\bar{x}=\frac{64+\cdots+68}{5}=65.6$$

2. 計算離差　　-1.6 -13.6　4.4　8.4　2.4

3. 計算平均絕對離差 $MAD=\frac{1.6+13.6+4.4+8.4+2.4}{5}=\frac{30.4}{5}=6.08$

練習：接續例 3-9，**計算六位女同學期中考成績的平均絕對離差。**

3.3.3 全距

全距(range, R)的定義為一組資料的最大值(Maximum)減去最小值(Minimum)，表示資料內部任何兩點的最大距離，自然成為一種表示資料分散程度的統計量數。全距的優點除了定義簡單易懂外，就是計算容易，在過去計算機尚未成熟的時期，有限資料下，全距是最容易計算的分散趨勢統計量。

例 3-21

接續例 3-9，計算五位男同學期中考成績的全距。

解：

1. 排序

52, 64, 68, 70, 74

2. 計算全距　　R=74-52=22

練習：接續例 3-9，**計算六位女同學期中考成績的全距。**

以上介紹的四種分散趨勢量數(變異數，標準差，平均絕對離差和全距)有一個共通的特點，就是都用到資料的最大值與最小值，如果一個統計量數使用到最大值或最小值，就會使得這個統計量數容易受到極端值的影響。這是很清楚的，一組資料若有極端值，那麼最大值與最小值必定也是極端值之一，所以，變異數，標準差，平均絕對離差都是容易受到極端值影響的統計量數。但其中以全具統計

量數受極端值最為嚴重，因為最大值與最小值之差就是代表資料兩端極端值的距離。

全距雖然容易受到極端值影響卻不靈敏，無法反映最小值與最大值間資料的變化。相對的，異數，標準差和平均絕對離差考慮到所有數值，雖容易受到極端值影響，但也能反映資料內所有觀察值的變化。

例 3- 22

分別計算下列三組資料的平均值，中位數，變異數，標準差，平均絕對離差和全距。

A 組　1, 2, 3, 4, 5, 6, 7, 8, 9, 10

B 組　1, 2, 3, 4, 5, 6, 7, 8, 9, 100

C 組　1, 2, 2, 3, 3, 8, 8, 9, 9, 10

解：

1. 計算平均值

$$\bar{x}_A = \frac{1+2+3+4+5+6+7+8+9+10}{10} = 5.5$$

$$\bar{x}_B = \frac{1+2+3+4+5+6+7+8+9+100}{10} = 14.5$$

$$\bar{x}_C = \frac{1+2+2+3+3+8+8+9+9+100}{10} = 5.5$$

2. 計算中位數

$$Med_A = 5.5$$

$$Med_B = 5.5$$

$$Med_C = 5.5$$

3. 計算樣本變異數

$$s_A^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2+10^2-10\times5.5^2}{9} = \frac{55}{6}$$

$$s_B^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2+100^2-10\times14.5^2}{9} = \frac{5455}{6}$$

$$s_C^2 = \frac{1^2+2^2+2^2+3^2+3^2+8^2+8^2+9^2+9^2+10^2-10\times5.5^2}{9} = \frac{73}{6}$$

4. 計算樣本標準差

$$s_A = \sqrt{s_A^2} = \sqrt{55/6} = 3.028$$
$$s_B = \sqrt{s_B^2} = \sqrt{5455/6} = 30.152$$
$$s_C = \sqrt{s_C^2} = \sqrt{73/6} = 3.567$$

5. 計算平均絕對離差

絕對離差值

A 組 4.5 3.5 2.5 1.5 0.5 0.5 1.5 2.5 3.5 4.5

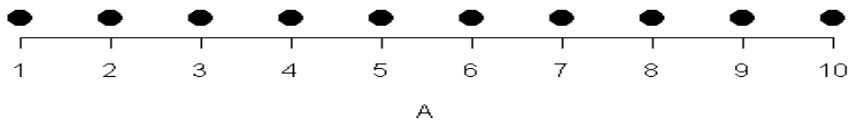
B 組 13.5 12.5 11.5 10.5 9.5 8.5 7.5 6.5 5.5 85.5

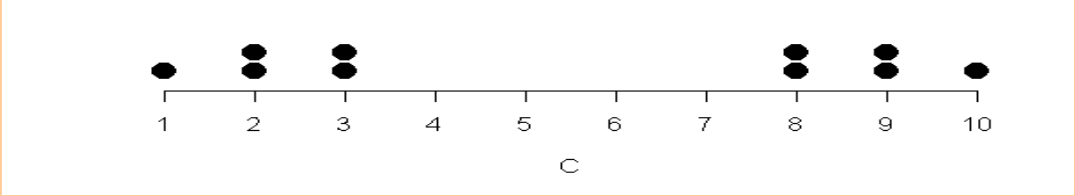
C 組 4.5 3.5 3.5 2.5 2.5 2.5 2.5 3.5 3.5 4.5

$$MAD_A = \frac{4.5+3.5+\cdots+3.5+4.5}{10} = 2.5$$
$$MAD_B = \frac{13.5+12.5+\cdots+5.5+85.5}{10} = 17.1$$
$$MAD_C = \frac{4.5+3.5+\cdots+3.5+4.5}{10} = 3.3$$

6. 計算全距

$$R_A = 10-1 = 9$$
$$R_B = 100-1 = 99$$
$$R_C = 10-1 = 9$$





A, B, C 三組資料的點圖

比較：

(1) **比較 A 和 B 圖：**扣除 B 圖的最大值，兩個圖形的分散程度相近，然四種統計量數都顯示兩組資料的分散程度明顯不同，尤以全距統計量最為明顯。

(2) **比較 A 和 C 圖：**兩個圖形的全距相同，然 C 圖觀察值較集中於兩側，表示資料較為分散，然全距統計量無法呈現彼此的差異。

3.3.4 四分位距

四分位距(inter-quartile range, IQR)也是一種衡量資料分散程度的統計量數，其定義為第 3 個四分位數減去的 1 個四分位數，即

$$IQR = Q_3 - Q_1$$

四分位距是一種排除最大值與最小值的統計量數，所以，它相對之前介紹的分散趨勢量數，具有穩定性的優點，不容易受到極端值的影響。

例 3- 23

計算表 2-1，晶圓厚度的四分位距。

因為 $Q_1 = 1460$，$Q_3 = 1517$，所以，$IQR = 1517 - 1460 = 57$。

例 3- 24

接續例 3- 22, 分別計算下列三組資料的四分位距。

A 組　1, 2, 3, 4, 5, 6, 7, 8, 9, 10

B 組　1, 2, 3, 4, 5, 6, 7, 8, 9, 100

C 組　1, 2, 2, 3, 3, 8, 8, 9, 9, 10

第一個四分位數 Q_1：

因為 $Q_1 = P_{25}$，$i = 25$，位置指標 $k = \frac{25}{100} \times 10 = 2.5$，所以，$Q_1 = x_{(3)}$。

第三個四分位數 Q_3：

因為 $Q_3 = P_{75}$，$i = 75$，位置指標 $k = \frac{75}{100} \times 10 = 7.5$，所以，$Q_3 = x_{(8)}$。

	Q_1	Q_3	IQR
A	3	8	5
B	3	8	5
C	2	9	7

四分位距(IQR)反映資料中心附近的分散程度，所以，A 和 B 的四分位距並無差異。此例說明四分位距不受極端值的影響，但受到內部資料改變的影響。

3.3.5 變異係數

試想我們要比較不同群體間的分散程度，例如比較國小一年學生的體重間的分散程度和國小六年級學生的體重間的分散程度，何者較為嚴重？在國民財富上，分散程度就是一種貧富差異的程度，如何比較哪一個國家的貧富差距比較嚴重？利用上述分散趨勢統計量數來比較是否和理呢？

卡爾皮爾生(Karl Pearson, 1858-1936)在觀看過很多實際測量的數據後，發現分散程度的大小與集中趨勢量數有一定的關聯性，通常呈現正比的關係。國小六年級學生的平均體重大過國小一年級學生的平均值，而他的變異程度也是比較大的。在有錢的國家，他們民眾間最有錢與最沒錢間的差距也是大。所以，卡爾皮爾生提出變異係數（coefficient of variation）來比較不同群體間的分散程度。

變異係數 $CV=\frac{\text{標準差}}{\text{平均值}}\times 100\%$ 式 3- 13

例 3- 25

請利用變異係數比較國小一年級與六年級學生體重的分散程度。

國小一年級 22.9 22.4 24.9 23.8 23.5 $\bar{x}=23.5, s=0.95$

國小六年級 44.7 37.9 41.4 36.7 33.4 $\bar{x}=38.82, s=4.36$

解：

國小一年級 $CV=\frac{0.95}{23.5}\times 100\%=4\%$

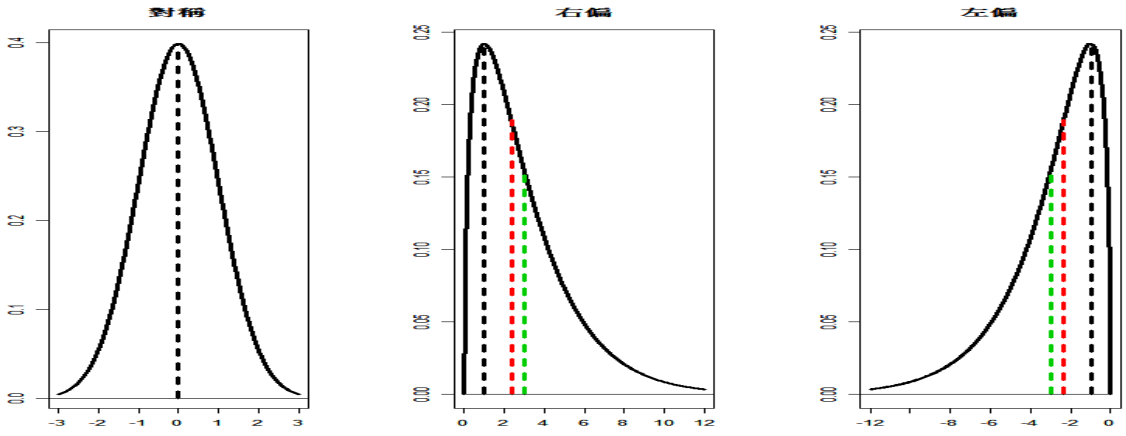
國小六年級 $CV=\frac{4.36}{38.82}\times 100\%=10.9\%$

由變異係數比較，我們認為國小六年級學生的差異較大。

例 3- 26

請至圖書館或上網搜尋，比較各國貧富差距的方法。

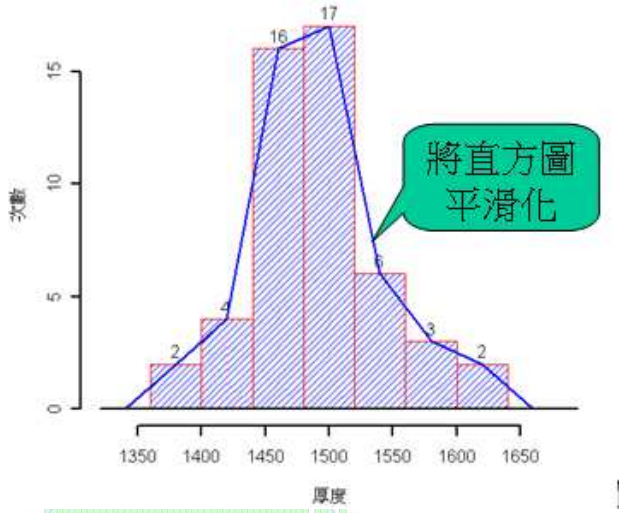
3.4 資料的偏態



就上圖來說，資料的偏態分成三種，右偏、左偏和對稱。一個單峰的分配，分配曲線左右兩端逐漸遞減，我們稱為左尾(left-tailed)與右尾(right-tailed)。若

是左尾與右尾的形狀如同鏡射，如上圖右，我們稱為對稱分配。若是右尾拖長，我們稱為右偏分配（right-skewed distribution），相對若是左偏拖長，則稱為左偏分配。上圖中，綠線是平均值，紅線是中位數，黑線是眾數。

實務上，我們可繪製資料的直方圖(histogram)再平滑化，利用平滑後的分配曲線辨識資料的偏態。然而，有時利用視覺方法通常很難論斷，也無法提供形狀偏向的程度。所以，我們介紹利用統計量判別資料的偏態。



當分配對稱時，三條線重疊，如右圖；當分配右偏時，三者關係為平均值>中位數>眾數；當分配左偏時，三者關係為平均值<中位數<眾數。我們注意到，不管是哪一種情形，中位數都是位於平均數與眾數之間，所以，卡爾皮爾生(Karl Pearson)提出皮爾生偏態係數衡量資料的編態

皮爾生偏態係數

母體 $$KP_s = \frac{3(\mu - Med)}{\sigma}$$ 式 3- 14

樣本 $$KP_s = \frac{3(\bar{x} - Med)}{s}$$

由分子的平均值與中位數相差可以判斷資料的偏態，若分子為正表示平均數大於中位數，資料為右偏，若分子為負正表示平均數小於中位數，資料為左偏。

例 3- 27

接續例 3- 9，5 位男同學的成績為 64，52，70，74，68，計算男同學的偏態係數。

因為 $\bar{x}=65.6$，$Med=68$，變異數為

$$s^2=\frac{64^2+52^2+70^2+74^2+68^2-5\times 65.6^2}{4}=\frac{21800-21516.8}{4}=70.8$$

標準差 $s=\sqrt{s^2}\sqrt{70.8}=8.41$，所以

$$KP_s=\frac{3(65.6-68)}{8.41}=-0.856$$

例 3- 28

請問表格 2-1 的數據的偏態。

因為 $\bar{x}=1491.02$，$Med=1483.5$，$s=50.41$，所以

$$KP_s=\frac{3(1491.02-1483.5)}{50.41}=0.448$$

3-5 謝比雪夫不等式與經驗法則

統計學的核心就是企圖瞭解一個群體的分配。如果知道這件事，我們就可以清楚知道在每一個範圍內的個體佔群體的比例，也知道每一個個體在群體中的相對位置。譬如，假設你的入學考成績為 300 分，這個成績在所有考生的排名會影響到你是否可以進入理想學校；又如，一個產品都有一定的規格，產品的量測值落在規格界線內，該產品就會被判定成良品。因此，當我們知道產品量測值的分配，我們就可以知道該產品的良率是多少，我們會賣出多少比例不良品，可以評估不良品對公司的影響，加強檢驗或是提供備品給顧客換貨。

這一節我們要介紹兩種利用平均數與標準差衡量資料落在給定範圍內的比例。

謝比雪夫不等式

(1) 不論資料的分配為何，至少有 $1-\frac{1}{k^2}$ 的資料落在以平均數為中心的 k 個標準差

之內。或者說

(2)不論資料的分配為何，最多有$\frac{1}{k^2}$的資料落在以平均數為中心的k個標準差之外。

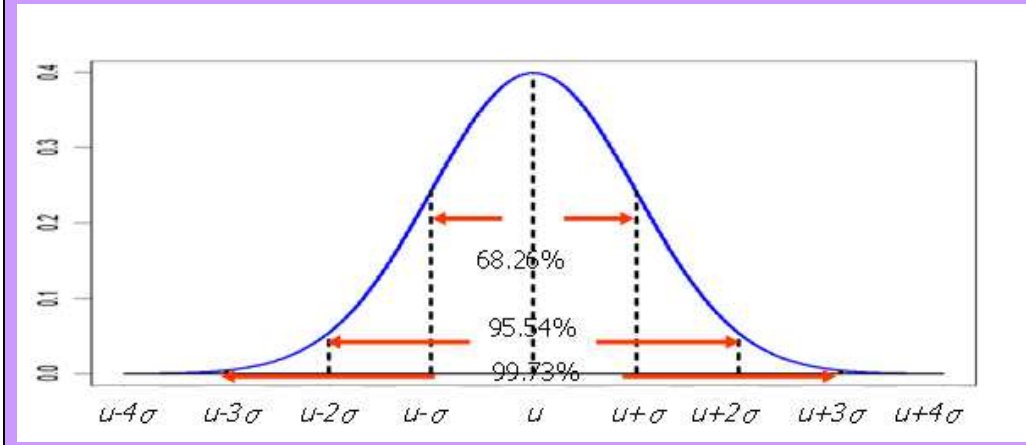
例 3- 29

本年級男新生有 1000 人，假設身高的平均值為 168 分，標準差為 4 公分。請問：

(1) 身高介於 160~176 公分的男同學至少有多少人？

(2) 身高介於 158~178 公分的男同學至少有多少人？

經驗法則(68-95-99.73 法則)



若資料呈現鐘形分配，則

(1)約有 68.26%的資料落在 1 個標準差之內$u\pm\sigma$。

(2)約有 95.54%的資料落在 2 個標準差之內$u\pm 2\sigma$。

(3)約有 99.73%的資料落在 3 個標準差之內$u\pm 3\sigma$。

謝比雪夫不等式是一般性的結果，這個性質是對任何資料都是成立的，但有時也顯得過於保少。當資料分配接近鐘形，經驗法則可以得到不錯的近似效果。

例 3- 30

接續例 3- 29，假設男同學身高分布為鐘形，請問：

(1) 身高介於 160~176 公分的男同學約有多少人？

(2) 身高介於 158~178 公分的男同學約有多少人？

例 3- 31

晶圓片厚度的平均值為 1491.02, 標準差為 50.4。

1379 1385 **1412** 1425 1427 1428 **1448** 1449 1451 1455 1459 1460 1460 1461 1461

1462 1465 1469 1470 1470 1472 1474 1482 1483 1483 1484 1488 1495 1500 1500

1501 1505 1509 1512 1512 1513 1516 1517 1520 1533 1543 1546 1547 **1548** 1554

1566 1568 **1569** 1601 1614

1 個標準差之內 $\bar{x}\pm s=(1440.62,1541.42)$，實際落在範圍內的比例為 $\frac{38}{50}=0.76$。

2 個標準差之內 $\bar{x}\pm 2s=(1390.22,1591.82)$，實際落在範圍內的比例為 $\frac{46}{50}=0.92$

3 個標準差之內 $\bar{x}\pm 3s=(1339.82,1642.22)$，實際落在範圍內的比例為 1.00。

3-6 探索式資料分析法

探索式資料分析法(Exploratory Date Analysis, EDA)是以簡單的圖形或結合一些統計量數，能夠對資料分配有粗略的瞭解。所謂探索，並不限定以下要介紹的幾種方法，任何一種以發現資料特性的方法都可以稱做探索式資料分析法。探索本意就是企圖發現我們本來不知的特性，論證某些現象是否依然存在等等，而非去呈現我們已知的現象。所採用的各種技術，希望可以達到

1. 對資料有全面性的瞭解
2. 能夠發現過去不知道的資料結構

3. 有無極端值或異常值存在

4. 檢定一些現象是否仍然保有

5. 提供模型參數簡化的資訊

以下我們介紹兩種探索式資料分析法

3.6.1 枝葉圖

簡單說枝葉圖(stem-and-leaf display)就是以保留原有數據繪製的直方圖。通常，繪製的方法是以數據最小的位數為葉，其他位數為枝。例如，繪製同學考試成績的枝葉圖，考試成績通常是兩位數，所以，我們以個位數葉，十位數為枝。我們用以下的例子來說明完成枝葉圖的過程。

例 3- 32

以下是 50 位同學考試的成績，請繪製同學考試成績的枝葉圖

74 64 58 69 70 72 75 74 68 75 84 71 85 77 57 74 86 73 62 69 **54** 77 67 59 61

69 59 58 84 70 57 76 **93** 67 72 64 61 60 74 78 58 65 80 72 61 65 75 69 54 60

1. 求得數據的最大與最小值。

最大值為 93，最小值為 54

2. 所需要枝為 5, 6, 7, 8, 9

5	
6	
7	
8	
9	

4. 繪製枝葉圖(沒有排序)，把個位數依序填入對應的枝。

5	874998784
6	49829719741051590
7	402545174370624825
8	45640
9	3

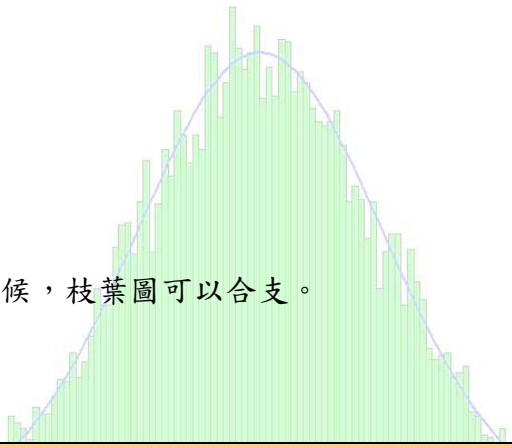
4. 將葉排序

```
5 | 447788899
6 | 00111244557789999
7 | 001222344445556778
8 | 04456
9 | 3
```

當樹枝內的葉太多的時候，枝葉圖可以分支為 2。上枝為{0,1,2,3,4}，下枝為{5,6,7,8,9}。可得到

```
5 | 44
5 | 7788899
6 | 00111244
6 | 557789999
7 | 00122234444
7 | 5556778
8 | 044
8 | 56
9 | 3
```

當樹枝內的葉太少的時候，枝葉圖可以合支。



例 3- 33

利用 R 程式繪製晶圓片厚度的枝葉圖。

```
142 | 0289
144 | 0239934678
146 | 44469
148 | 6736899
150 | 1283337
152 | 1123666456
154 | 31
156 | 345
158 | 1
160 | 5
```

圖 3- 1：莖葉圖 (stem-and-leaf display)

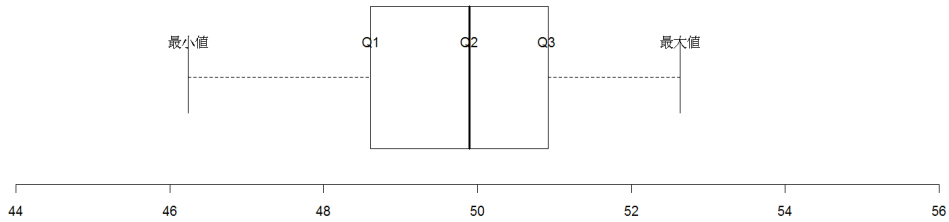
這是一個合枝枝葉圖。

R 程式語法：

```
stem(x)
```

3.6.2 盒鬚圖

盒鬚圖(box-and-wisher plot)又可以稱為盒型圖，箱型圖等。它是最普遍的一種探索式資料分析工具，利用五個統計量：最小值、最大值和Q_1,Q_2,Q_3，繪製一個有兩條長鬚的箱型圖。



盒鬚圖提供了以下幾點資訊：

(1). 資料的集中趨勢

(2). 資料的分散趨勢

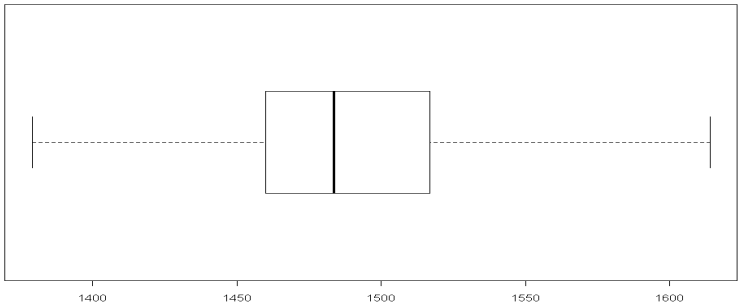
(3). 資料的形狀

換言之，一個盒鬚圖就可以傳達本章講授的內容，可見其在應用上的重要性。

例 3- 34

請繪製表格 2-1，50 片晶圓厚度的盒形圖。

根據 例 3- 15，Q_1=1460,Q_2=1483.5,Q_3=1517 。



此盒形圖顯示資料略有右偏。

盒型圖主要應用在不同群體間分配的比較。

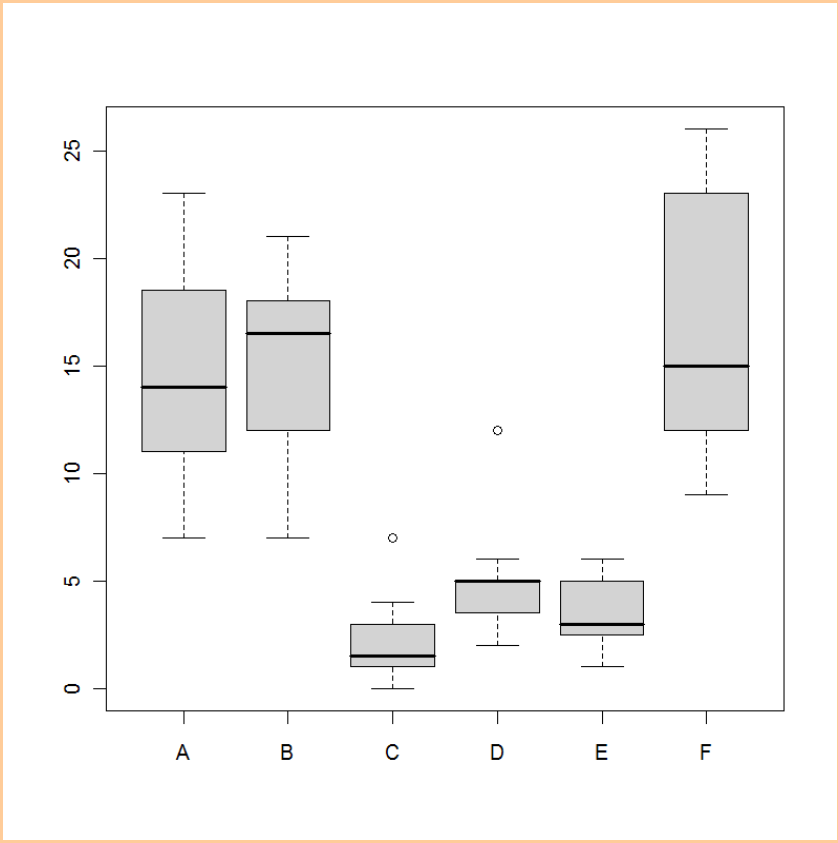
例 3- 35

下圖是 6 個廠房每日產品不良率的盒形圖(R 程式中的例子)，將它們的不良盒形圖放在一張圖上，除了可以清楚顯示每一個廠房的品質狀況，主要還能比較它們間的差異性。

(1) 這六個盒形圖差異頗大，表示各廠房的品質狀況並不相同。

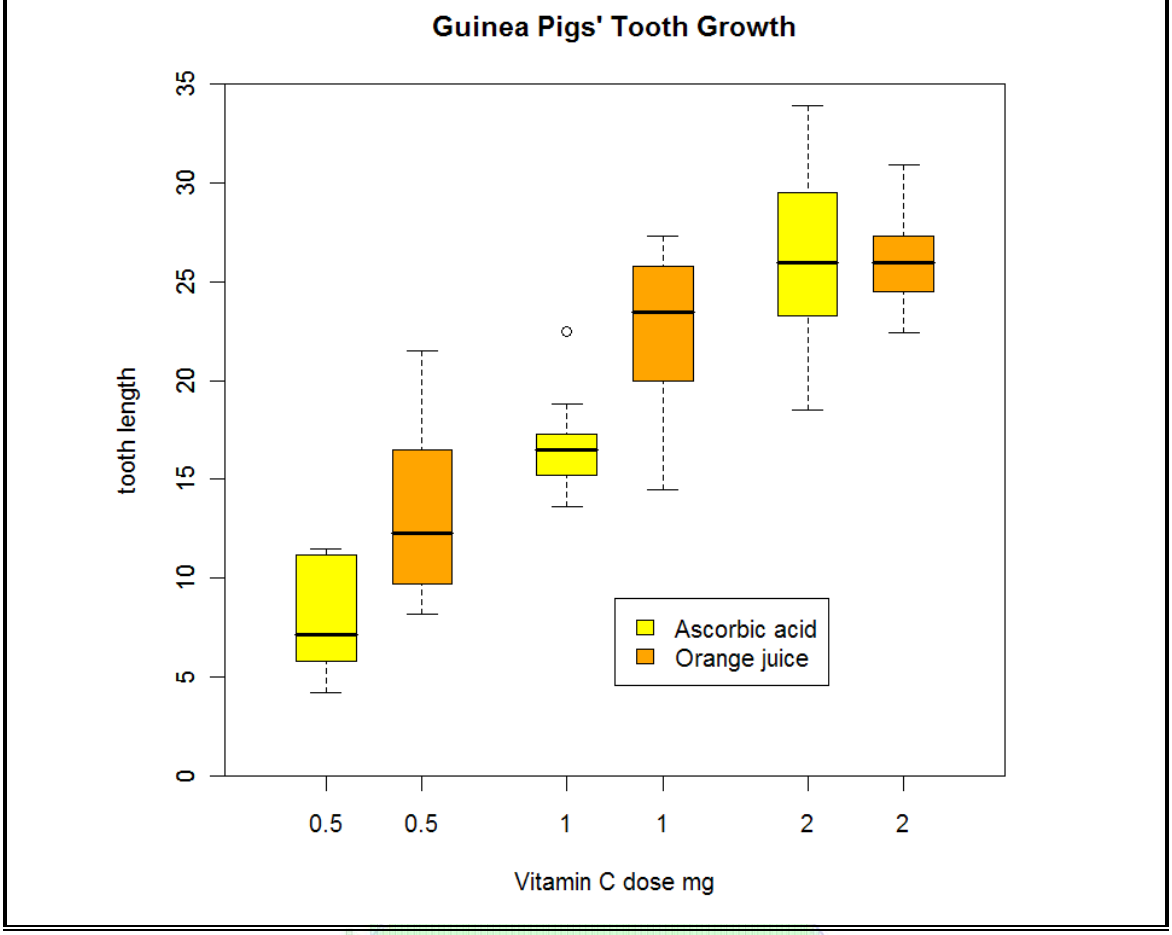
(2) 盒子內的粗線為中位數，可粗略將六個廠房分成兩組。以 C, D, E 廠房的不良率較低，而相對而言，A, B, F 廠房不良率較高。

(3) 這三家不良率較高的廠房，盒子的長度比較長，表是良率變化較其他三家為大。



例 3- 36

豬隻牙齒長度與維他命 C 攝取量的關係(R 程式中的例子)



習題

1. 計算以下樣本統計量

29 31 35 35 40 41 44 45

(a)Mean _____

(b)Median_____

(c)Mode_____

(d)Q_1 _____

(e)Q_3_____

(f)P_{20} _____

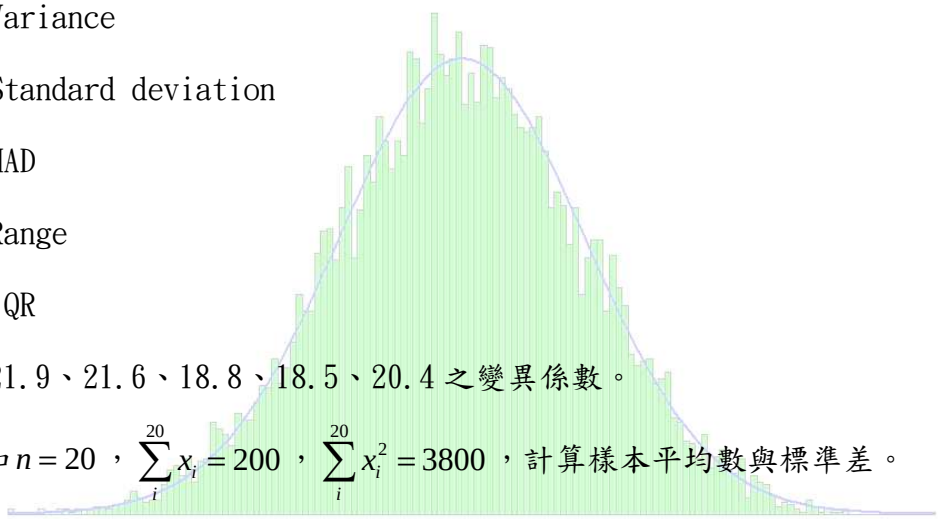
(g)Variance

(h)Standard deviation

(i)MAD

(j)Range

(k)IQR



2. 求 21.9、21.6、18.8、18.5、20.4 之變異係數。

3. 已知 $n=20$，$\sum_{i}^{20} x_i = 200$，$\sum_{i}^{20} x_i^2 = 3800$，計算樣本平均數與標準差。

4. 已知 $n=10$，$\sum_{i}^{10} x_i = 150$，$\sum_{i}^{10} x_i^2 = 2500$，計算樣本平均數與標準差。

5. 一組樣本 201 202 203 208 196，請求出平均數離差。請驗證平均數離差和為零。

6. 調查 40 位同學每日的上網時間(小時)，資料如下：

0.9 1.0 1.0 1.0 1.0 1.1 1.2 1.2 1.3 1.4

1.4 1.4 1.4 1.4 1.5 1.5 1.6 1.8 1.8 1.8

1.9 2.0 2.0 2.0 2.0 2.0 2.1 2.1 2.3 2.4

2.5 2.6 2.6 2.7 2.7 2.7 2.8 3.1 3.3 3.4

(a) 請繪製枝葉圖。

(b) 請求 Q_1, Q_2, Q_3。

(c) 請繪製盒形圖。

7. 若某種類型的墊圈厚度μ=1.95㎜，σ=0.12㎜，在1000個墊圈中應有多少個厚度在1.80與2.10㎜之間。

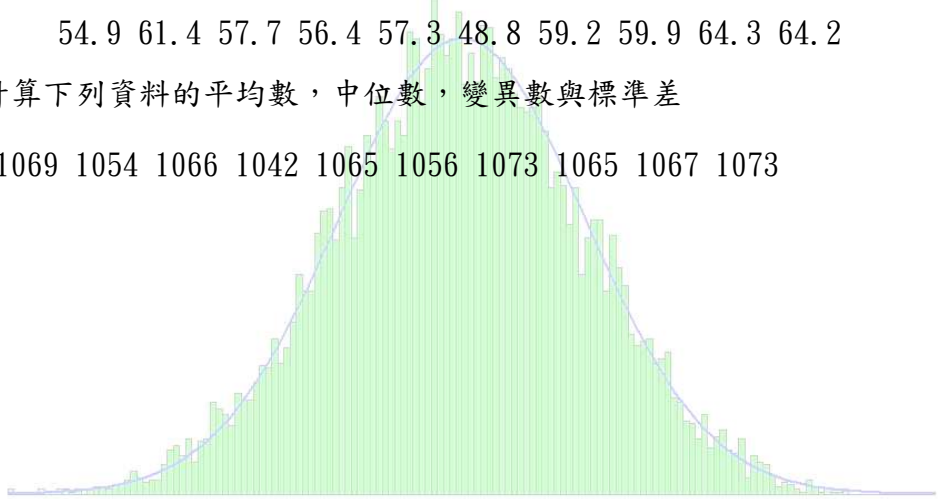
8. 假設甲乙兩班人數為50和60，平均身高為167和169，請問兩班同學的平均身高為多少。

9. 請計算下列資料的三個四分位數，並繪製盒形圖。

46.4 66.9 65.1 46.7 64.9 62.0 62.2 61.8 60.2 48.8
80.0 56.5 52.7 71.4 42.2 79.0 55.0 47.2 57.0 51.9
55.0 61.6 48.5 54.9 61.2 63.7 61.6 65.5 61.9 42.8
67.2 50.1 49.6 52.8 39.3 60.7 53.8 68.7 52.9 66.2
54.9 61.4 57.7 56.4 57.3 48.8 59.2 59.9 64.3 64.2

10. 請計算下列資料的平均數，中位數，變異數與標準差

1069 1054 1066 1042 1065 1056 1073 1065 1067 1073



重點摘錄