

3 統計量

>

Question: 一個資料分配的特徵值可以從哪幾個方面進行描述？

使用統計表、統計圖將一群散亂的資料加以彙整呈現，便能迅速瞭解資料的特徵和形狀。但要進一步分析資料、瞭解資料的數值特徵時，就需要利用一些統計量數來測度與描述資料的性質。描述資料特性的量數，主要有(1)集中趨勢量數；(2)位置統計量數；(3)分散趨勢量數；以及(4)其他統計量數。

學習目標：

1. 學習集中趨勢統計量
2. 學習位置統計量數
3. 學習分散趨勢統計量
4. 其他統計量數

Answer 一組資料分配的特徵可以從三方面進行描述：一是資料的水準（集中趨勢、位置度量），反應資料的數值大小；二是資料的差異，反應各資料間的離散程度；三是分配的形狀，反應資料分配的偏度和峰度。

3-1 集中趨勢量數

集中趨勢量數(measures of central tendency)，係用來衡量資料的中心位置，反映該組資料的共同趨勢與集中位置。一般較常用的集中趨勢量數有平均數、中位數、眾數等。

Question: 說明平均數、中位數和眾數的特點和應用場合。

3-1-1. 平均數

平均數(mean)：一組資料相加後，除以資料的個數所得到的結果。

- 衡量資料中心位置最重要的測量數。
- 使用到每一個觀察值，包括極端值。

1. 計算方式（算術平均數）

- 母體平均數

假設母體資料總數為 N ，以 X_1, X_2, \dots, X_N 表示，則母體平均數以 $\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \dots + X_N}{N}$ 表示。

- 樣本平均數

假設樣本資料總數為 n ，以 x_1, x_2, \dots, x_n 表示，則樣本平均數 $\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$

例題 3.1

國軍某部隊舉行射擊訓練，共有 10 位士兵參加，每人射擊 6 發，擊中靶心的次數如下：

4 3 4 2 3 4 6 5 4 6

(1) 求擊中靶心次數的樣本平均數 \bar{x} ？

(2) 求擊中靶心比例的樣本平均數？

解 (1) 全部資料的加總為 $\sum_{i=1}^{10} x_i = 4 + 3 + 4 + 2 + 3 + 4 + 6 + 5 + 4 + 6 = 41$ 。

共有 10 位士兵參加，表示 $n = 10$ 。

因此擊中靶心次數的樣本平均數 $\bar{x} = 41/10 = 4.1$ (次)

(2) 計算「擊中靶心比例」的樣本平均數，需先把這 10 位士兵擊中靶心次數換算成擊中靶心的比例，也就是資料全部除以 6，可得

$\frac{4}{6} \quad \frac{3}{6} \quad \frac{4}{6} \quad \frac{2}{6} \quad \frac{3}{6} \quad \frac{4}{6} \quad \frac{6}{6} \quad \frac{5}{6} \quad \frac{4}{6} \quad \frac{6}{6}$

將其全部相加再除以資料總數就是擊中靶心比例的樣本平均數

$$\left\{ \frac{4}{6} + \frac{3}{6} + \frac{4}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{6}{6} + \frac{5}{6} + \frac{4}{6} + \frac{6}{6} \right\} / 10 = \frac{4.1}{6} = 0.68$$



例題 3.2

某高中甲班學生 30 人之體重（單位：公斤）如下：

37	39	45	40	41	42	38	37	46	43
38	39	37	47	46	47	43	38	37	40
44	45	45	46	43	39	37	38	41	42

(1) 以簡單隨機抽樣法隨機選取其中五位學生（樣本）並量測其體重如下：

45	46	40	41	38
----	----	----	----	----

求其平均數？

(2) 計算全班（母體）之平均體重？

解 (1) 樣本平均數 \bar{x} （5 位學生的平均體重）

$$\bar{x} = \frac{45+46+40+41+38}{5} = 42 \text{ 公斤}$$

(2) 母體平均數 μ （全班的平均體重）

$$\mu = \sum_{i=1}^{30} x_i / 30 = \{37+39+\cdots+42\} / 30 = 1240 / 30 = 41.3 \text{ 公斤}$$

- 每次抽樣的樣本不一定會相同，因此統計量的值（樣本平均數）會隨著每次抽樣結果的不同而不同。
- 母體永遠不會改變，因此母體參數的值（母體平均數）永遠固定。
- 樣本平均數和母體平均數並不相同。此差距是因為抽樣所致，稱為抽樣誤差(sampling error)。

例題 3.3

隨機抽取 10 位民眾，得到月薪資料（單位：萬元）如下：

2	6	4	5	4	5	5	6	7	100
---	---	---	---	---	---	---	---	---	-----

請計算其平均月薪，並討論此平均月薪是否為一個好的代表值。

解

$$\bar{x} = \frac{2+6+4+5+4+5+5+6+7+100}{10} = \frac{144}{10} = 14.4 \text{ (萬元)}$$

平均月薪為 14.4 萬元。

這組資料中，多數人的月薪都在 4 萬到 6 萬之間，即多數人的月薪都無法到達 14.4 萬，但是卻因為有一個月薪 100 萬的資料，導致平均月薪被拉高為 14.4 萬。在這樣的狀況下，以平均數做為資料中心點和代表值並不太適當。



(2) 平均數的特性

- 優點：

- 平均數的代表性容易被接受
- 平均數只有一個，具唯一性。
- 資料中的每一數值均被列入計算。

- 缺點：

- 平均數容易受極端值的影響。若是資料中出現極端大，或是極端小的極端值 (extreme values) 資料，就會導致平均數沒有集中趨勢統計量所該有的特性，造成我們使用上會誤解資料的資訊。

3. 加權平均數

假設權數分為 w_1, w_2, \dots, w_n ，且 $\sum w_i = 1$ ，樣本加權平均數為 $\bar{x}_w = w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n$ 。

例題 3.4

小明各科成績與每週上課時數如下：

科 目	成績 (分)	每週上課時數
國 文	80	2
英 文	70	2
會計學	75	3
微積分	60	3
統計學	85	3

試求算 (1) 加權平均數？

(2) 平均數？

解 (1) 加權平均數

$$\bar{x}_w = \frac{2}{13} \times 80 + \frac{2}{13} \times 70 + \frac{3}{13} \times 75 + \frac{3}{13} \times 60 + \frac{3}{13} \times 85 = \frac{960}{13} = 73.85 \text{ (分)}$$

(2) 平均數

$$\bar{x} = \frac{80 + 70 + 75 + 60 + 85}{5} = \frac{370}{5} = 74 \text{ (分)}$$



3-1-2. 中位數

將資料依照數值大小的順序，由小到大排序後，可以找出排在某個位置上的數值，用該數值代表資水準的高低。

中位數 (median)：居於數列中間位置的那個數值，以 M_e 表示。

- 中位數用一個點 (位置)，將資料分成2個等分。
 - 全部資料分成兩部分，每部分包含50%的資料。
 - 一半的資料比中位數大。
 - 一半的資料比中位數小。

(1) 計算方式

假設一組資料 x_1, x_2, \dots, x_n 有 n 個數值，先將資料由小到大排序為 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ ，然後確定中位數的位置，最後確定中位數的具體數值。

- n 為奇數時，中位數在第 $\frac{n+1}{2}$ 項，即 $x_{(\frac{n+1}{2})}$ 。
- n 為偶數時，中位數為第 $\frac{n}{2}$ 項與第 $\frac{n}{2} + 1$ 項的平均數，即 $\frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]$ 。

例題 3.5

隨機詢問 5 個家庭的年度國民所得（單位：萬元）如下：

100	95	80	120	110
-----	----	----	-----	-----

求中位數 M_e 。

解 先將資料由小至大排列如下：

80	95	100	110	120
----	----	-----	-----	-----

選擇居中的數據， $n=5$ 為奇數，中位數在第 $\frac{5+1}{2}=3$ 項。

中位數為 $M_e = 100$ （萬元）。

例題 3.6

隨機詢問 6 個家庭的年度國民所得（單位：萬元）如下：

75	90	80	95	100	110
----	----	----	----	-----	-----

求中位數 M_e 。

解 先將資料依小至大排列如下：

75	80	90	95	100	110
----	----	----	----	-----	-----

$n=6$ 為偶數，中位數為第 $\frac{6}{2}=3$ 項與第 $\frac{6}{2}+1=4$ 項的平均數

中位數為 $M_e = \frac{90+95}{2} = 92.5$ （萬元）



(2) 中位數的特性

- 優點：
 - 不受極端值的影響。
 - 若資料中存在有極端值時，中位數較能代表資料的中心位置。
 - 中位數一定存在，且具唯一性。

- 缺點：
 - 只考慮居中位置的數值，忽略其他數值的大小。
 - 求中位數時，需先將資料由小至大排序，當資料量大時，排序工作的困難度就會提高。

3-1-3. 眾數

眾數(mode)：一組資料中出現次數最多的那一個數值，以 M_0 表示。

~ 眾數的特性

- 長條圖或是直方圖最高的柱子出現的地方
- 眾數可能有許多個，也可能一個都沒有。
 - 若出現一個眾數，則為單峰，此為資料的中心位置。
 - 若是雙峰或多峰，則眾數不代表中央位置。
 - 眾數比平均數和中位數少用。
- 不受極端值影響
- 也適用於類別變數

例題 3.7

國軍某部隊舉行射擊訓練，共有 10 位士兵參加，每人射擊 6 發，擊中靶心的次數如下：

4 3 5 2 3 5 6 5 5 6

試問眾數 M_o 為何？

解 次數分配表如下：

擊中靶心次數	次數 (人數)
2	1
3	2
4	1
5	4
6	2
合計	10

出現最多次的是「擊中靶心次數為 5」的士兵，在 10 人中有 4 人。

眾數 $M_o = 5$

例題 3.8

請求出下列各小題的眾數：

- (1) 2, 3, 4, 5, 6, 7
- (2) 2, 2, 3, 3, 5, 5, 5, 6, 7
- (3) 2, 2, 3, 3, 5, 5, 5, 6, 70
- (4) 1, 2, 2, 3, 4, 4, 5

- 解**
- (1) 每個數值只出現一次，故沒眾數。
 - (2) 「5」出現 3 次，為最多次數，故眾數為 5。
 - (3) 「5」出現 3 次，為最多次數，故眾數為 5。
 - (4) 「2」和「4」皆出現 2 次，為最多次數，故眾數為 2 和 4。



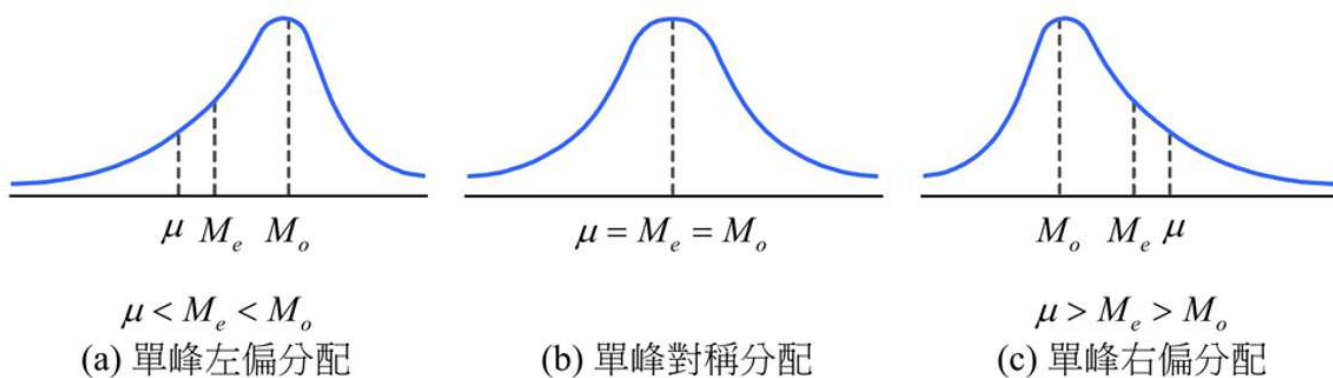
3-1-4 平均數、中位數及眾數之關係與比較

=> 在統計學的範圍中，集中趨勢統計量常用的有三種：

- 平均數是最常見的一種集中趨勢統計量。
- 中位數是與資料順序有關的集中趨勢統計量。
- 眾數是與資料的次數分配有關的集中趨勢統計量。

在衡量中央位置時，平均數通常是第一個選擇，但當有極端值時，中位數較能代表資料的中心位置。

=> 左偏、對稱與右偏分配



- 單峰對稱分配
 - 曲線為對稱分配，此時平均數、中位數與眾數，呈現三點合一的情形。
- 單峰左偏配
 - 曲線為左偏分配，此時平均數最小，中位數則介於平均數與眾數之間
- 單峰右偏配
 - 曲線為右偏分配，此時平均數為最大，且呈現與左偏分配相反之位置分布

3-2. 位置統計量數

=> 位置統計量，描述在資料群中相對位置的統計量。

- 與中位數類似的位置統計量，還有四分位數(quartile)、十分位數(decile)和百分位數(percentile)等。他們分別是用3個點、9個點和99個點，將資料分成4等分、10等分和100等分。
- 例如統計學期中考，張三考了45分，在班上50人中排名40。請問比張三成績低的有多少百分比？如果老師要當百分之20，請問張三會不會被當掉？

3-2-1. 百分位數

1. Def: 第 p 百分位數(p -th percentile)

將資料按順序由小到大排列後，若至少有 $p\%$ 的觀測值在某一數值之下，至少有 $(100-0)\%$ 的觀測值在該數值之上，則此數值稱為該組資料的第 p 百分位數。

2. 計算步驟

Step1: 先將資料由小到大順序排列。

Step2: 計算 p 百分位數的位置指標： $i = \frac{n \cdot p}{100}$ 。

- p 為欲探討的百分位數。
- n 為觀察值個數。

Step3: 位置指標 i 是否為整數。

- 若 i 為非整數，則無條件進入到整數位的位置，其對應的資料值，即為第 p 百分位數；
- 若 i 為整數，則第 p 百分位數為位置 i 與 $i + 1$ 資料值之平均值。

例題 3.9 (承例題 2.3)

民國 108 年底佳佳經紀公司 50 名女性模特兒體重（單位：公斤）如下：

45	40	46	41	44	43	48	42	45	45
42	41	46	45	45	40	45	50	44	42
50	45	44	40	40	37	46	42	45	43
43	40	38	40	44	45	46	46	39	51
44	38	39	39	43	46	40	46	38	44

請計算：

- (1) 此組資料之 50 百分位數。
- (2) 此組資料之 60 百分位數。
- (3) 此組資料之 75 百分位數。

解 先將資料排序，結果為：

37	40	43	45	46
38	40	43	45	46
38	40	43	45	46
38	40	43	45	46
39	41	44	45	46
39	41	44	45	46
39	42	44	45	48
40	42	44	45	50
40	42	44	45	50
40	42	44	46	51

資料共有 50 個數值 $\Rightarrow n = 50$ 。

1. 50 百分位數

$$\frac{np}{100} = \frac{50 \times 50}{100} = 25$$

位置指標為整數，所以 50 百分位數為第 25 項和第 26 項數值的平均數，即 $\frac{44+44}{2} = 44$ 。

2. 60 百分位數

$$\frac{np}{100} = \frac{50 \times 60}{100} = 30$$

位置指標為整數，所以 60 百分位數為第 30 項和第 31 項數值的平均數，即 $\frac{44+45}{2} = 44.5$ 。

3. 75 百分位數

$$\frac{np}{100} = \frac{50 \times 75}{100} = 37.5$$

位置指標為非整數，所以無條件進位，75百分位數為第38項的數值，即45。

3-2-2 四分位數

四分位數(quartile): 特殊的百分位數

1. Def: 將一組資料排序之後，用三個點將全部資料分割成四等份，每個部分包含四分之一的資料。在25%、50%與75%這三個點位置上的數值，就是四分位數。
2. 四分位數的特性
 - 四分位數有三個
 - 第1個四分位數： $Q_1 = P_{25}$
 - 第2個四分位數： $Q_2 = P_{50}$
 - 第3個四分位數： $Q_3 = P_{75}$
 - 第*i*個四分位數記為 $Q_i, i = 1, 2, 3$
 - 至少有 $i/4$ 的觀察值小於等於該數值
 - 至少有 $(4 - i) / 4$ 的觀察值大於等於該數值。

例題 3.10 (承例題 3.9)

民國 108 年底佳佳經紀公司 50 名女性模特兒體重，請計算：

- (1) 此組資料之第 1 四分位數 Q_1 。
- (2) 此組資料之第 2 四分位數 Q_2 。
- (3) 此組資料之第 3 四分位數 Q_3 。

解 資料共有 50 個數值 $\Rightarrow n = 50$

- (1) 第 1 四分位數 $Q_1 \Rightarrow 25$ 百分位數 $\Rightarrow p = 25$

$$\frac{np}{100} = \frac{50 \times 25}{100} = 12.5 \text{ (不整數)} \Rightarrow \text{無條件進入到整數位：13} \Rightarrow \text{找出第}$$

13 項的數值

第 1 四分位數 $Q_1 = 40$

Excel 2016 函式：`=PERCENTILE.INC (A1:E10,0.25)`

- (2) 第 2 四分位數 $Q_2 \Rightarrow 50$ 百分位數 $\Rightarrow p = 50$

$$\frac{np}{100} = \frac{50 \times 50}{100} = 25 \text{ (整數)} \Rightarrow \text{找出第 25 項和第 26 項的數值}$$

$$\text{第 2 四分位數 } Q_2 = \frac{44 + 44}{2} = 44$$

Excel 2016 函式：`=PERCENTILE.INC (A1:E10,0.5)`

- (3) 第 3 四分位數 $Q_3 \Rightarrow 75$ 百分位數 $\Rightarrow p = 75$

$$\frac{np}{100} = \frac{50 \times 75}{100} = 37.5 \text{ (不整數)} \Rightarrow \text{無條件進入到整數位：38} \Rightarrow \text{找出第}$$

38 項的數值

第 3 四分位數 $Q_3 = 45$

Excel 2016 函式：`=PERCENTILE.INC (A1:E10,0.75)`



3-3. 分散趨勢量數

1. 分散趨勢量數的目的

- 以一個簡單的數字，來表示一群資料數值分散的程度。
- 用以區別集中趨勢統計量代表性的強弱。

例題 3.11

假設有以下三組資料，每組資料各有 5 個數值，三組資料如下：

第一組	5	5	5	5	5
第二組	3	4	5	6	7
第三組	1	3	5	7	9

試計算各組的平均數，並討論各組資料的分散狀況。

解 第一組平均數 $= \frac{5+5+5+5+5}{5} = 5$

第二組平均數 $= \frac{3+4+5+6+7}{5} = 5$

第三組平均數 $= \frac{1+3+5+7+9}{5} = 5$

各組的平均數都是 5，但是第一組 5 個資料都是 5，完全集中於同一個數值；第二組稍微分散，每個數值都相差 1；第三組最分散，每個數值都相差 2。雖然平均數都是 5，但是仍然出現資料分散狀況不同的情況。這說明了只看平均數（或是集中趨勢統計量），會有一些資料的特性被忽視了。



2. 離散趨勢量數是用來衡量各觀測值間之分散情況。

- 離散統計量的數值越大，表示資料越分散。
- 重要的離散統計量有全距與四分位距、變異數與標準差等。

3-3-1. 全距與四分位距

1. 全距(range)

1. Def: 資料中數值最大者與最小者之差距

$$Range = Max - Min = x_{(n)} - x_{(1)} = 100\text{百分位數} - 0\text{百分位數}。$$

2. 全距的性質

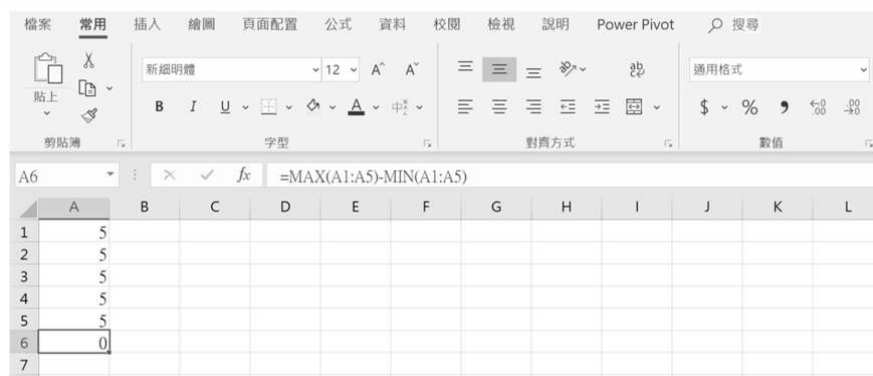
- 全距容易受極端值影響。

例題 3.12 (承例題 3.11)

請計算各組資料之全距。

解 (1) 第一組資料全部都是 5，由小到大排序後為 5，5，5，5，5。

因此第一組資料的全距 $R_1 = 5 - 5 = 0$



The screenshot shows the Excel 2016 interface. The formula bar at the top displays the formula `=MAX(A1:A5)-MIN(A1:A5)` for cell A6. The worksheet grid shows columns A through L and rows 1 through 7. Column A contains the values 5, 5, 5, 5, 5 in rows 1 through 5, and 0 in row 6. The other cells in the grid are empty.

Excel 2016 函式：`=MAX(A1:A5)-MIN(A1:A5)`

(2) 第二組資料由小到大排序後為 3，4，5，6，7。

因此第二組資料的全距 $R_2 = 7 - 3 = 4$

(3) 第三組資料由小到大排序後為 1，3，5，7，9。

因此第三組資料的全距 $R_3 = 9 - 1 = 8$



2. 四分位距 (inter-quartile range)

1. Def: 資料經由小至大排序後，第三個四分位數和第一個四分位數的差距。

$$IQR = Q_3 - Q_1 = 75\text{百分位數} - 25\text{百分位數}。$$

- 資料的第三個四分位數 Q_3 ，稱為上四分位數 (upper-quartile)。
- 資料的第一個四分位數 Q_1 ，稱為下四分位數 (lower-quartile)。

2. 四分位距特性

- 相較於全距，四分位距不易受極端值影響。
- 四分位距常與中位數搭配使用
 - 以中位數表示集中趨勢
 - 以四分位距表示離散趨勢

例題 3.13 (承例題 3.11)

請計算各組資料之四分位距。

- 解** (1) 第一組資料全部都是 5，因此 $Q_3 = Q_1 = 5$ ， $IQR_1 = Q_3 - Q_1 = 5 - 5 = 0$ 。
- (2) 先計算第二組資料的 25 百分位數和 75 百分位數
- (a) 第二組資料的 25 百分位數 (Q_1^*)： $5 \times 25/100 = 1.25$ 不是整數，因此無條件進位為 2， $Q_1^* = x_{(2)} = 4$ 。
- (b) 第二組資料的 75 百分位數 (Q_3^*)： $5 \times 75/100 = 3.75$ 不是整數，因此無條件進位為 4， $Q_3^* = x_{(4)} = 6$ 。
- (c) $IQR_2 = Q_3^* - Q_1^* = 6 - 4 = 2$ 。
- (3) 先計算第三組資料的 25 百分位數和 75 百分位數
- (a) 第三組資料的 25 百分位數 (Q_1^+)： $5 \times 25/100 = 1.25$ 不是整數，因此無條件進位為 2， $Q_1^+ = x_{(2)} = 3$ 。
- (b) 第三組資料的 75 百分位數 (Q_3^+)： $5 \times 75/100 = 3.75$ 不是整數，因此無條件進位為 4， $Q_3^+ = x_{(4)} = 7$ 。
- (c) $IQR_3 = Q_3^+ - Q_1^+ = 7 - 3 = 4$ 。

3-3-2 (box plot)

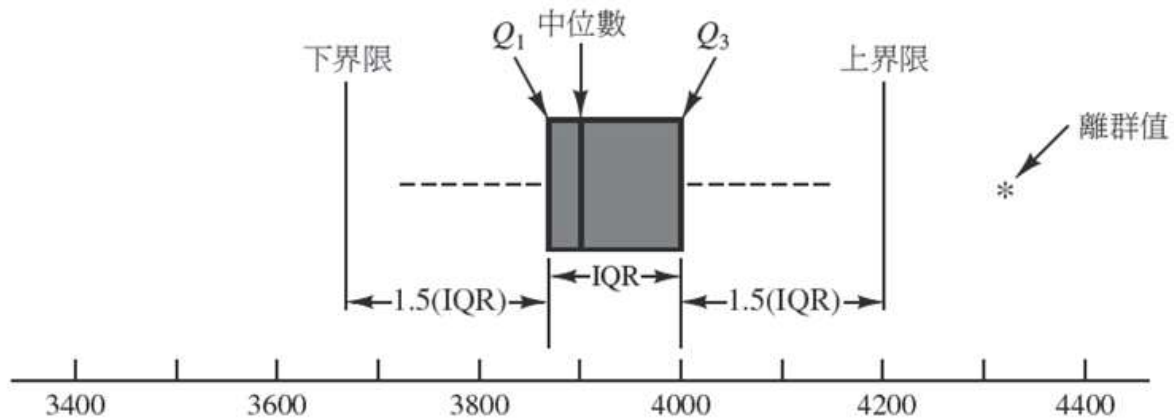
箱形圖是一種圖形，與以下這五種數字有關：

- 數據集的最小值。
 - Q_1
 - 中位數
 - Q_3
- 數據集的最大值。

這五個數字叫做數據集的五數摘要 (five-number summary)。

圖 3.6

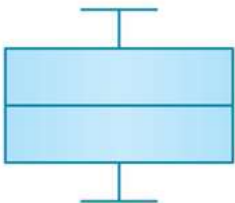
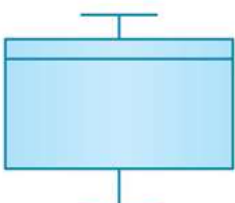
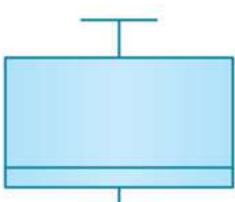
每月起薪資料的箱形圖以及上、下界限



~ 建構盒形圖

- **步驟 1:** 求出數據集的五數摘要，也就是找到最小值、 Q_1 、中位數、 Q_3 和最大值。
- **步驟 2:** 假設完成後的盒形圖擺在水平線上。
 - 畫一段 x 軸
 - 包含最小值和最大值
 - 畫一個盒子
 - 四分位距的 Q_1 和 Q_3 為邊界
 - 代表資料集中趨勢之情況
 - 盒中垂直線為中位數的位置
 - 頭尾垂直線
 - 最大值與最小值。
 - 連接盒子和頭尾。
 - 離群值(outlier)
 - 小於 $Q_1 - 1.5 * IQR$
 - 大於 $Q_3 + 1.5 * IQR$
 - 表示資料離散的狀況

表 3.1 直式箱型圖與資料的分配的關係

箱型圖	資料的分配
	對稱分配
	左偏分配 (負偏)
	右偏分配 (正偏)

3-3-3. 變異數與標準差

1. 變異數(variance)：離差平方的平均數

- 每一個觀察值與平均數之間的差距，稱為離差， $x_i - \bar{X}$ 。
- 將所有離差平方加總，再求其平均數，可得離差平方的平均數，也就是平均平方離差，即為變異數。
 - 母體變異數： $\sigma^2 = \frac{\sum_{i=1}^N (x - \mu)^2}{N}$
 - 樣本變異數： $S^2 = \frac{\sum_{i=1}^n (x - \bar{x})^2}{n-1} = \frac{(\sum_{i=1}^n x^2 - n\bar{X})^2}{n-1}$

2. 標準差(standard deviation)

- 變異數開根號的正值
 - 母體標準差 $\sigma = \sqrt{\sigma^2}$
 - 樣本標準差 $S = \sqrt{S^2}$
- 標準差使用單位與原來的使用單位相同
- 觀察值離平均數越遠，標準差越大

~ 變異數與標準差會用到每一筆資料。

例題 3.14

小南排骨飯在全台灣有共十家門市，某天營業額為 12、9、20、6、8、4、7、9、15、6 (萬) 元，試問：

- (1) 母體變異數 σ^2 是多少？
- (2) 母體標準差 σ 是多少？
- (3) 若在十家門市中，抽取五家門市，其營業額分別為：

12、8、4、15、6 (萬) 元

樣本變異數 s^2 是多少？

- (4) 續 (3)，樣本標準差 s 是多少？

解 (1) 先計算母體平均數 μ

$$\mu = \frac{12+9+20+6+8+4+7+9+15+6}{10} = 9.6 \text{ (萬) 元}$$

母體變異數 σ^2 為

$$\sigma^2 = \frac{1}{10}[(12-9.6)^2 + (9-9.6)^2 + \cdots + (15-9.6)^2 + (6-9.6)^2] = 21.04$$

3-4. 機率計算 - 平均數與標準差的應用

集中趨勢量中的平均數與離勢量數中的標準差，這兩種量數除自身能提供訊息之外，結合應用這兩種量數，可進一步提供資料分佈的描述分析方法。

3-4-1. 變異係數

Question: 為什麼要計算變異係數？

例：我們有班上50位同學的身高與體重資料，請問班上同學是身高還是體重的差異性（離散程度）較大？

例：我們有班上50位同學的學測英文與數學成績，請問班上同學是英文還是數學的差異性（離散程度）較大？

(1) 變異係數(coefficient of variation, CV): 用來衡量資料的相對分散程度。

當兩組資料的觀測值相差較大，或者是資料使用單位不同時，可用變異係數做比較。

2. 計算公式

- 母體變異係數 $CV = \frac{\text{標準差}\sigma}{\text{平均數}\mu} * 100$
- 樣本變異係數 $CV = \frac{\text{標準差}s}{\text{平均數}\bar{x}} * 100$

(3) 衡量不同組資料的相對分散程度

- 變異係數是個百分比，表示標準差相對於平均數大小的百分比。

- CV 越小，標準差相對於平均數很小，資料集中在平均數附近
- CV 越大，標準差相對於平均數很大，資料比較分散。

例題 3.18

甲班 50 名學生平均身高 μ_1 為 160 公分，標準差 σ_1 為 10 公分；而平均體重 μ_2 為 50 公斤，標準差 σ_2 為 4.2 公斤，試以變異係數比較身高與體重的資料散布狀態。

解 身高與體重單位不同，比較時必須先計算變異係數。

$$\text{身高之變異係數 } CV_1 = \frac{\sigma_1}{\mu_1} \times 100\% = \frac{10}{160} \times 100\% = 6.25\%$$

$$\text{體重之變異係數 } CV_2 = \frac{\sigma_2}{\mu_2} \times 100\% = \frac{4.2}{50} \times 100\% = 8.4\%$$

$CV_1 < CV_2$ ，身高的變異係數較小，身高資料比體重資料更集中。



隨堂練習

例題 3.19

高雄市一歲以下嬰兒平均身高 μ_1 為 50 公分，標準差 σ_1 為 5 公分。二十歲以上之成年人平均身高 μ_2 為 167 公分，標準差 σ_2 為 7 公分。試比較兩組資料之變異係數，何者較小？這兩組資料中，哪組資料的散布狀態較為集中？

解 兩者身高之單位相同，但由於平均身高差異太大，不宜直接比較，故應計算其變異係數。

$$\text{嬰兒之變異係數 } CV_1 = \frac{\sigma_1}{\mu_1} = \frac{5}{50} \times 100\% = 10\%$$

$$\text{成人之變異係數 } CV_2 = \frac{\sigma_2}{\mu_2} = \frac{7}{167} \times 100\% = 4.19\%$$

成人身高資料之變異係數比較小，因此成人身高資料的散布狀態較為集中。



3-4-2. Z分數 (標準化分數)

標準化分數 (standard score)：把一組資料轉換成具有平均數為0、標準差為1新資料。

1. 標準化分數通常用大寫的Z或是小寫的z表示，又稱為z分數(z-score)。

- 表示某一個數值與平均數的差距是幾倍的標準差。

(2)計算公式

- 標準化母體資料： $z = \frac{x_i - \mu}{\sigma}$
- 標準化樣本資料： $z = \frac{x_i - \bar{x}}{s}$

3. 功用

- 去掉原始資料的單位，可以進行不同單位的資料比較。
 - 瞭解某筆資料值在整個資料集中的相對位置。
 - Z 分數的絕對值越大，代表此數值離平均數越遠
 - Z 分數的絕對值越小，代表此數值離平均數越近
4. 標準化分數只是將原始資料進行線性變換，它並沒有改變某個數值在該組資料中的位置，也沒有改變該組資料分配的形狀。

例題 3.20

甲班 50 名學生統計學期中考的平均成績 μ_1 為 62 分，標準差 σ_1 為 5 分；而微積分期中考的平均成績 μ_2 為 68 分，標準差 σ_2 為 2 分。這次期中考，小雨的統計學成績 x_1 是 60 分，微積分成績 x_2 為 64 分。試求小雨的統計學與微積分成績的 z 分數，並說明小明在班上是統計學還是微積分的成績較佳？

解 因為統計學、微積分是不同的兩個學科，要進行比較，需先化為 z 分數

$$\text{統計學 } z \text{ 分數} = \frac{x_1 - \mu_1}{\sigma_1} = \frac{60 - 62}{5} = -0.4$$

$$\text{微積分 } z \text{ 分數} = \frac{x_2 - \mu_2}{\sigma_2} = \frac{64 - 68}{2} = -2$$

小雨統計學的 z 分數為 -0.4 ，表示其統計學成績值低於班上平均成績 0.4 個標準差，而微積分的 z 分數為 -2 ，表示其微積分成績低於於班上平均成績 2 個標準差。故小雨的統計學比微積分在班上較佳。



例題 3.21

某大學以三科學測成績決定錄取。今需要從甲、乙兩位考生中選出一位，試討論應該選擇哪一位？

	考生級分		全體考生	
	甲生	乙生	平均數	標準差
國文	10	9	9	1.5
英文	8	10	8	2
數學	11	10	7	2.5

解 首先計算甲、乙兩位考生各科目的 z 分數

	考生級分		全體考生		z 分數	
	甲生	乙生	μ	σ	甲生	乙生
國文	10	9	9	1.5	0.67	0
英文	8	10	8	2	0	1
數學	11	10	7	2.5	1.6	1.2
總分	29	29			2.27	2.2

甲、乙兩位考生的原始分數，三科級分總和相同，因此無法由此決定哪一位考生勝出。但比較 z 分數時，甲考生的國文和數學都稍優於乙考生，僅英文比乙考生差。若錄取標準注重英文，則應該錄取乙考生；若注重國文或數學能力，就應該錄取甲考生；若以此三科 z 分數的總和，則甲考生險勝乙考生。

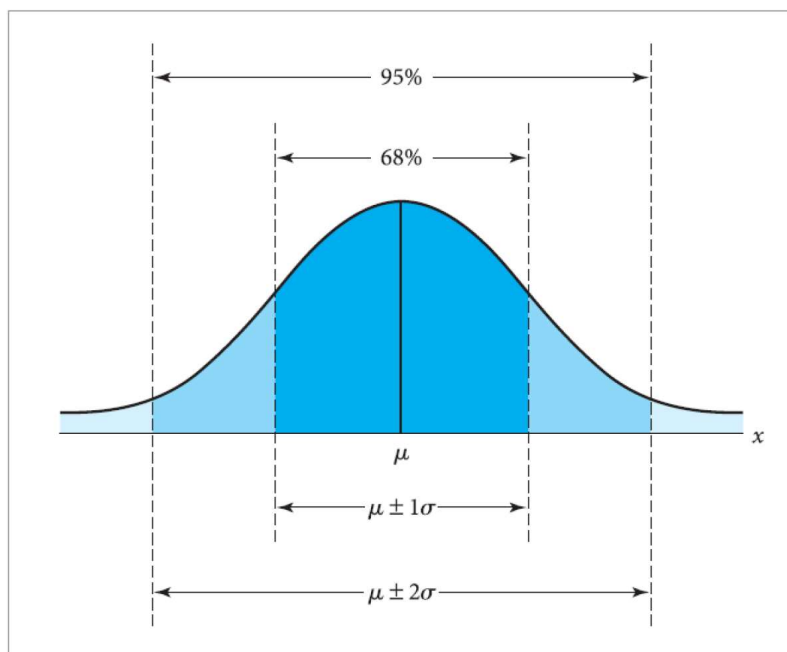


3-4-3. 機率的計算

(1) 經驗法則(empirical rule)

Question: 一群資料中到底有多少百分比會落在平均數加減幾個標準差的範圍內？

- 如果資料的直方圖看起來像「鐘型」，我們計算出平均數 \bar{X} 及標準差 S 之後，就可以大約得知一組資料落在平均數加減一、二、三個標準差範圍內之比例，稱為經驗法則。結果整理如下：
 - 約有68%之資料會落在平均數加減一個標準差的範圍內，即 $(\mu - \sigma, \mu + \sigma)$ 之區間內。
 - 約有95%之資料會落在平均數加減兩個標準差的範圍內，即 $(\mu - 2\sigma, \mu + 2\sigma)$ 之區間內。
 - 約有99.7%之資料會落在平均數加減三個標準差的範圍內，即 $(\mu - 3\sigma, \mu + 3\sigma)$ 之區間內。



例題 3.22

假設企管系有一百名學生，統計學期中考的平均成績為 66 分，標準差為 6 分，試以經驗法則求算下列問題：

- (1) 統計學期中考成績在 60~72 分的學生約有多少人？
- (2) 統計學期中考成績在 54~78 分的學生約有多少人？
- (3) 統計學期中考成績在 48~84 分的學生約有多少人？

解

(1) $(60, 72) = (66 - 6, 66 + 6) \Rightarrow$ 平均數加減 1 個標準差之距離

故約有 68% 的比例落於 60~72 分範圍內，約 $100 \text{ 名} \times 68\% = 68 \text{ 名}$ 學生。

(2) $(54, 78) = (66 - 2 \times 6, 66 + 2 \times 6) \Rightarrow$ 平均數加減 2 個標準差之距離

故約有 95% 的比例落於 54~78 分範圍內，約 $100 \text{ 名} \times 95\% = 95 \text{ 名}$ 學生。

(3) $(48, 84) = (66 - 3 \times 6, 66 + 3 \times 6) \Rightarrow$ 平均數加減 3 個標準差之距離

故約有 99.7% 的比例落於 48~84 分範圍內，約 $100 \text{ 名} \times 99.7\% = 99 \text{ 名}$ 學生。



(2) 謝比雪夫不等式 (Chebyshev's inequality)

如果資料的直方圖看起來不像「鐘型」，或者是在沒有任何假設或前提的情況下，給一個不用複雜計算、比較不精確的參考數值時，我們可以使用謝比雪夫不等式。

定理：一群資料至少有 $(1 - \frac{1}{k^2}) * 100$ 的比例會落在平均數附近 k 個標準差的區間範圍內。換句話說，落於 $(\mu - k\sigma, \mu + k\sigma)$ 之區間內，至少有 $(1 - \frac{1}{k^2}) * 100$ 比例的資料，其中 k 必須大於 1。

k	區間	落於該區間內觀測值的比例
1	$(\bar{x} - s, \bar{x} + s)$	至少為 0 (至少 0%)
2	$(\bar{x} - 2s, \bar{x} + 2s)$	至少為 $\frac{3}{4}$ (至少 75%)
2.5	$(\bar{x} - 2.5s, \bar{x} + 2.5s)$	至少為 $\frac{21}{25}$ (至少 84%)
3	$(\bar{x} - 3s, \bar{x} + 3s)$	至少為 $\frac{8}{9}$ (至少 89%)

例題 3.23

假設企管系有一百名學生，統計學期中考的平均成績為 66 分，標準差為 6 分，試以謝比雪夫定理求算下列問題：

- (1) 統計學期中考成績在 54~78 分的學生約有多少人？
- (2) 統計學期中考成績在 48~84 分的學生約有多少人？

解

(1) $(54, 78) = (66 - 2 \times 6, 66 + 2 \times 6) \Rightarrow k = 2$

至少有 $\left(1 - \frac{1}{k^2}\right) \times 100\% = \left(1 - \frac{1}{2^2}\right) \times 100\% = 75\%$ 在此範圍內

故至少 75 名同學，統計學期中考成績在 54~78 分。

(2) $(48, 84) = (66 - 3 \times 6, 66 + 3 \times 6) \Rightarrow k = 3$

至少有 $\left(1 - \frac{1}{3^2}\right) \times 100\% \approx 88.9\%$ 在此範圍內

故至少有 $100 \text{ 名} \times 88.9\% = 88.9 \text{ 名}$ 同學，無條件進位為 89 人，統計學期中考成績在 48~84 分。



例題 3.24

今年學測數學成績平均級分為 8 分，標準差 2 分，試求：

- (1) 成績在 4 分與 12 分之間約有多少比例的人？
- (2) 成績在 6 分與 14 分之間約有多少比例的人？

假設成績適合於 (a) 任何分配，或 (b) 鐘形分配方式，試分別求算之。

解 (a) 假設成績為任何分配，依謝比雪夫定理：

(1) $(4, 12) = (8 - 2 \times 2, 8 + 2 \times 2) \Rightarrow k = 2 \Rightarrow$ 至少 75% 的考生成績在 4~12 分之間。

(2) $(6, 14) = (8 - 3 \times 2, 8 + 3 \times 2) \Rightarrow k = 3 \Rightarrow$ 至少 88.9% 的考生成績在 6~14 分之間。

(b) 假設成績呈鐘形分配，依經驗法則：

(1) $(4, 12) = (8 - 2 \times 2, 8 + 2 \times 2) \Rightarrow k = 2 \Rightarrow$ 大約 95% 的考生成績在 4~12 分之間。

(2) $(6, 14) = (8 - 3 \times 2, 8 + 3 \times 2) \Rightarrow k = 3 \Rightarrow$ 大約 99.7% 的考生成績在 6~14 分之間。



Summary: 經驗法則與Chebyshev定理的比較

區間	Chebyshev 定理	經驗法則
$(\bar{x} - s, \bar{x} + s)$	無意義	約 68 %
$(\bar{x} - 2s, \bar{x} + 2s)$	至少 75%	約 95 %
$(\bar{x} - 3s, \bar{x} + 3s)$	至少 89%	約 99.7 %