03 敘述統計(II)——統計量數

3.1 集中趨勢量數
3.2 差異量數
3.3 平均數與標準差的一些應用
3.4 分組資料之各種統計量數的計算
3.5 探索性資料分析
3.6 Excel範例

著作權所有 侵害必究

第2章討論了各種統計圖與表，有助於我們了解一組資料的特性。但為了獲得具體的描述與比較分析資料，則須進一步地求出資料之位置或集中趨勢的量數，以及資料的差異量數。這些量數即稱為**統計量數**(statistical measures)，它是代表一組資料之特質的具體數字。由於統計量數是數字性的，故可做一般的數學分析。

描述統計資料之特性的統計量數，主要有下列二項：

1. 集中趨勢量數：陳示一組統計資料的代表值，亦即其中心位置之所在。
2. 差異量數：陳示一組統計資料的各個觀測值對於集中趨勢數值或某標準值之差異。

除了上述二種統計量數外，尚有偏態量數與峰度量數，但限於篇幅，本書不擬介紹。本章將依序介紹集中趨勢量數與差異量數，這些量數的計算皆只針對未分組的資料。至於已分組資料之各種量數的計算，將在第3.4節專門討論。最後，第3.5節將討論一種新近發展出來的統計資料之分析方法，此即探索性資料分析。

華泰文化
HWA TAI PUBLISHING

3.1 集中趨勢量數

集中趨勢量數(measures of central tendency)亦簡稱為**集中量數**，它代表一組資料中，各個個體的某種特性有共同的趨勢存在之量數，又因其可反映該組資料觀測值集中的位置，故又稱為**位置量數**(location measures)。

較常用的集中趨勢量數有平均數、加權平均數、中位數、眾數與百分位數；本節將一一介紹。

3.1.1 平均數

平均數(mean)可說是最重要的集中趨勢量數，它可作為一組資料的代表值。一般而言，平均數具有簡化作用、代表作用及比較作用等幾項功用。

所謂簡化作用是指平均數能以一簡約的數字，概示一組資料分配的特徵。換句話說，平均數能簡化一組資料的全部數值而為一數值。代表作用是指平均數能代表一組資料的平均水準，此乃因為平均數是一組資料的中心數值，故以它來代表整個資料分配最為恰當。

至於比較作用則是由平均數之上述的二種功用衍生而來的，蓋平均數簡化所有的數值後，以該數值代表整組資料的平均水準，於是便可針對兩群或兩群以上的資料分配之間做比較。某一資料分配與另一資料分配之比較，唯有以具有代表整組資料之平均數來做比較才有意義。例如以平均成績而言，甲班的平均分數高於乙班，則可知道甲班的平均水準優於乙班。

此外，平均數依其平均方式的不同而有不同的種類，如**算術平均數**(arithmetic mean)、**幾何平均數**(geometric mean)與**調和平均數**(harmonic mean)，此三者各有其優缺點與適用場合。其中算術平均數乃是最為常用的平均數，本書主要即在介紹算術平均數（往後若未特別說明，則均以平均數代替算術平均數）。

平均數是集中趨勢量數中最簡單且最重要的量數，為最常被採用者。一組資料之全部數值相加的總和除以其所包含項數，由此所求得的即為平均數。在此所指的一組資料可能是母體，亦可能是樣本，但其觀念與計算方法皆相同，只是在符號上有些區別而已。以下所討論的各種量數，若有必要的時候，將特別指出母體或樣本的情況。

設一組資料有 n 個數值， $x_1, x_2, \cdots, x_n$，則其平均數為：

$$\text{樣本：}\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n} \tag{3-1}$$

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

$$\text{母體：}\mu=\frac{x_1+x_2\cdots+x_N}{N}=\frac{\sum x_i}{N}$$

上式中的$\overline{x}$與μ分別是樣本平均數與母體平均數的符號，而n與N分別代表樣本與母體之觀測值的總個數。

例 3.1

某次歌唱比賽冠軍總決賽，5位評審給2位參賽者的成績如下：

小榮：87, 89, 82, 84, 86
晴晴：86, 82, 85, 87, 84

請問那位參賽者獲得冠軍？

解 小榮：

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{5}x_i=\frac{1}{5}(87+89+82+84+86)$$

$$=\frac{428}{5}=85.6$$

晴晴：

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{5}x_i=\frac{1}{5}(86+82+85+87+84)$$

$$=\frac{424}{5}=84.8$$

比賽結果，小榮的平均成績為85.6，高於晴晴的平均成績84.8，所以此次比賽的冠軍是小榮。

由本例可知，平均數可用來比較兩組資料之平均水準，此乃平均數的功用。

例 3.2

設有三個班級甲、乙、丙，其學生人數分別為50、48、52人。某次統計學考試，此三個班級的平均成績分別為80、76、85。試求出此三個班級統計學之總平均成績。

解 依題意知：

$$n_1=50, n_2=48, n_3=52$$

$$\overline{x}_1=80, \overline{x}_2=76, \overline{x}_3=85$$

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

則總平均成績 $\overline{x}$ 為：

$$\overline{x} = \frac{50 \times 80 + 48 \times 76 + 52 \times 85}{50 + 48 + 52} = \frac{12{,}068}{150} = 80.45$$

例 3.3

洪姓果農今天採收阿露斯哈密瓜 55 顆，平均重量為 1,650 公克，客戶驗收時發現其中有 5 顆未達約定的品質標準而剔除，總重量 4,500 公克。請問客戶交貨的 50 顆哈密瓜平均重量為何？

解 令 $(n_1, \overline{x}_1)$ 及 $(n_2, \overline{x}_2)$ 分別表示 50 顆合格哈密瓜與 5 顆不合格哈密瓜之數量與平均重量。可得：

$$1,650 = \frac{n_1 \overline{x}_1 + n_2 \overline{x}_2}{n_1 + n_2} = \frac{50\overline{x}_1 + 4,500}{55}$$

由此可解得 $\overline{x}_1 = 1,725$；故 50 顆合格哈密瓜的平均重量為 1,725 公克。

3.1.2 加權平均數

前一節曾提及平均數是最常用的統計量數之一，其計算方法是將一組資料的所有數值加總，除以所包含的項數而得。此數值隱含一個假設，即所有數值或觀測值的重要性或權重都相同。雖然此法簡單易懂，但在某些情況下，為了反映個別觀測值的重要性，在計算平均數時須考量各觀測值的權重，即計算加權平均數 (weighted mean)。

樣本加權平均數的計算方式如下：

$$\overline{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3-2}$$

其中 w_i 為第 i 個觀測值的權重 (weight)，x_i 為第 i 個觀測值。若為母體資料，其計算方法與樣本加權平均數相同，只是符號不同而已。

比較 (3-1) 式與 (3-2) 式可知，若所有的 w_i 皆相同，則二式計算所得之結果完全相同；若所有 w_i 不相同，此二式所計算的結果則不同。換句話說，若所有的 w_i 皆為 1，則 (3-1) 式與 (3-2) 式相同。再從 (3-2) 式可知，當某個 x_i 的 w_i 較其他為高時，對平均數的影響會較大。例如，一般大學採加權平均數來計算學生學期總平均，以各科目的學分數為權數，因此學分數較高的科目，對學期總平均的影響較大。

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

例 3.4

亮亮參加某頂大會計系的甄選入學，第二階段各計分項目之權重與得分如下：

計分項目	權重	得分
學測成績	40%	90
審查資料	10%	85
面試	50%	84

請計算亮亮甄選第二階段所得到的分數。

解

計分項目	權重	得分	權重 × 得分
學測成績	40%	90	36.0
審查資料	10%	85	8.5
面試	50%	84	42.0
合計			86.5

著作權所有 侵害必究

加權平均數在日常生活中的應用相當廣泛，除了大學生的學期總平均外，還包括：大學外語系的入學成績會將英文加重計分，即將英文成績的權數設定比其他科目高，而計算加權平均成績；公司要招募員工時會設定相關的重要評估指標，並給予不同的權數，再針對每位應徵者進行評分，最後再以加權平均數作為錄取標準；公司在選擇工廠廠址時，亦會設定各項重要評估指標的權數，再針對每個可行的廠址進行評分，同樣會選擇加權平均數較高者。

3.1.3 中位數

當一組資料存在極端值時（即特別大或特別小的數值），其平均數的代表性受到影響，此時採用**中位數**(median)乃是較佳的量數。

一組 n 個數值的資料 $x_1, x_2, \cdots, x_n$，若按大小順序排列，則其中位數為位於中間位置的數值，亦即：當 n 為奇數時，第 $\frac{n+1}{2}$ 位置的數值為其中位數；當 n 為偶數時，$\frac{n}{2}$ 與 $\frac{n+2}{2}$ $\left(\text{或 } \frac{n}{2}+1\right)$ 位置之二數值的平均為其中位數。

概略地來說，中位數為將一組資料分為相等兩半的數值。換句話說，資料的 50% 位於中位數之下，而另外 50% 位於中位數之上。依此定義知，中位數著重於

華泰文化 HWA TAI PUBLISHING

位置，也就是說，求中位數時，即在於找出位於中央位置的那一個數值。中位數一般以符號 M_e 表示。

例 3.5

求下列二組資料之中位數：

Ⅰ：13, 20, 8, 15, 7
Ⅱ： 5, 10, 19, 23, 11, 15

解 先將資料按大小順序排列，然後找中間位置：

Ⅰ： 7, 8, 13, 15, 20
Ⅱ： 5, 10, 11, 15, 19, 23

$$\text{二數平均} = \frac{11+15}{2} = 13$$

Ⅰ、Ⅱ兩組資料之中位數均等於 13。

例 3.6

某個大學社團同屆畢業同學聚會，共 6 個同學參加，其去年的年收入分別為 40、45、250、48、60、75 萬元。試算其平均數與中位數，並加以比較說明。

解

$$\overline{x} = \frac{40+45+250+48+60+75}{6} = 86.3$$

計算中位數時，先將資料從小到大排列：

40, 45, 48, 60, 75, 250

因觀察值個數為偶數 6，故取第 3、4 位置兩個數值的平均數，即：

$$M_e = \frac{48+60}{2} = 54$$

平均數 86.3 萬元遠大於中位數 54 萬元，此乃因年收入最高（250 萬元）大幅提高了平均數。但觀察此 6 人去年的年收入，僅 1 人的年收入大於平均數 86.3 萬元。由此可知，此時以中位數 54 萬元做為資料的集中量數似乎較平均數為佳。

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

3.1.4 眾數

眾數(mode)亦為一種集中趨勢量數，一般均以符號 M_o 表示，其意義為：

一組資料中，出現次數最多的數值即為眾數。

由此定義知，眾數乃為一最集中且以位置為主的一個量數；諸如「一般的」、「慣例的」等用語，正是眾數的最佳註解。

例 3.7

試求出下列三組資料之眾數：

Ⅰ：15, 18, 20, 15, 15, 20, 25, 15
Ⅱ：10, 12, 10, 10, 8, 12, 12, 14
Ⅲ：2, 7, 5, 9, 16, 20, 8, 10

解 第Ⅰ組資料中，15 出現 4 次最多，故眾數為 15。
第Ⅱ組資料中，10 與 12 皆出現 3 次（最多），故眾數為 10 與 12。
第Ⅲ組資料中，各數值皆僅出現一次，故眾數不存在。
由此例子可知，**眾數可能不存在或者不只一個，因此眾數並不具有唯一存在性質**。

眾數特別適用於質的資料，而平均數與中位數則不然。例如某班級 40 位同學的血型資料如下：

血型	A	B	O	AB
人數	21	5	12	2

該班之血型以「A」最多，故血型資料的眾數為「A」。

3.1.5 百分位數

1. 百分位數的意義

依據中位數的定義，將資料按大小順序排列，則將資料分割成對等的兩半之數值即為中位數。將此觀念加以延伸，即可推出百分位數的意義：將資料依大小

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

順序排列，並取 99 個等分點（分割成 100 等分），則每一等分點皆為百分位數。百分位數具體的定義如下：

將資料按大小順序排列後，若至少有 $p\%$ 的觀測值位於某一數值以下，且至少有 $(100-p)\%$ 的觀測值位於該值以上，則此值稱為該組資料的第 p 個百分位數 (p-*th* percentile)。

第 k 個百分位數通常以 P_k 表示，例如第 32 個百分位數即為 P_{32}，其意義是：將一組資料（已排好大小順序）分布的範圍大小（全距的觀念）分割成相等的 100 等分，取第 32 個等分點，則其所在位置的數值即為 P_{32}。

百分位數的計算一般可依下列的步驟進行：

(1) 將資料依大小順序排列。
(2) 求出百分位數 (P_k) 所在位置的指標 (index)，設為 i，則 $i=\frac{k}{100}\times n$（n 表示觀測值的個數），$k=1, 2, \cdots, 99$。
(3) 若 i 為非整數，則 P_k 為下一個整數位置的數值，例如 $i=9.23$，則取第 10 個位置之數值為 P_k；另一方面，若 i 為整數，則取第 i 與 $i+1$ 位置的兩個數值之平均，即為所求的 P_k。

例 3.8

在某一十字路口測量噪音水準，記錄 50 個觀測值，由小而大依序排列，如表 3.1 所示。試求出 P_{25}、P_{30}、P_{50}、P_{75}。

表 3.1　噪音水準的觀測值（以分貝為單位）

52.0	55.9	56.7	59.4	60.2	61.0	62.1	63.8	65.7	67.9
54.4	55.9	56.8	59.4	60.3	61.4	62.6	64.0	66.2	68.2
54.5	56.2	57.2	59.5	60.5	61.7	62.7	64.6	66.8	68.9
55.7	56.4	57.6	59.8	60.6	61.8	63.1	64.8	67.0	69.4
55.8	56.4	58.9	60.0	60.8	62.0	63.6	64.9	67.1	77.1

解 (a) 求 P_{25}：

計算位置指標 $i=\frac{k}{100}\times n=\frac{25}{100}\times 50=12.5$，不為整數，故取第 13 個位置的數值，即 $P_{25}=57.2$。

(b) 求 P_{30}：

計算位置指標 $i=\frac{30}{100}\times 50=15$，為整數，故取第 15 與 16 兩個位置之數值的平均，即 $P_{30}=\frac{58.9+59.4}{2}=59.15$。

著作權所有 侵害必究

華泰文化 HWA TAI PUBLISHING

(c) 求 P_{50}：

計算位置指標 $i = \frac{50}{100} \times 50 = 25$，為整數，故取第 25 與 26 兩個位置之數值的平均，即 $P_{50} = \frac{60.8 + 61.0}{2} = 60.9$。

請注意，P_{50} 即為中位數 (M_e)；若依中位數的求法，亦可得出相同的結果。

(d) 求 P_{75}：

計算位置指標 $i = \frac{75}{100} \times 50 = 37.5$，不為整數，故取第 38 個位置的數值，即 $P_{75} = 64.6$。

2. 特殊的百分位數

百分位數的觀念是將一組資料分割成 100 個等分（每一等分皆含有相同個數的觀測值），取第 k 個等分點，即為 P_k；然這些等分點有些是較特殊的，例如：第 50 個等分點 (P_{50}) 乃將一組資料分割成兩等分，此即中位數的意義；第 25, 50, 75 個等分點（即 P_{25}, P_{50}, P_{75}）乃將一組資料分割成 4 個等分，此時這三個數值稱為**四分位數** (quartiles)；同理，第 $10, 20, 30, \cdots, 90$ 個等分點（即 $P_{10}, P_{20}, P_{30}, \cdots, P_{90}$）乃將一組資料分割成 10 個等分，此時這九個數值稱為**十分位數** (deciles)。

四分位數與十分位數是較常見的特殊百分位數，一般以 Q_i 與 D_i 表示。例如：Q_1 表示第一個四分位數，D_3 表示第三個十分位數，餘此類推。由此可知，四分位數與十分位數其觀念乃百分位數的推廣，因此其求法亦如同百分位數。例如，欲求 Q_1 只需求 P_{25}、欲求 D_3 只需求 P_{30} 等等。茲將百分位數、四分位數、中位數及十分位數之對照關係，列如表 3.2。

表 3.2　百分位數與中位數、四分位數、十分位數之對照

百分位數	P_{10}	P_{20}	P_{25}	P_{30}	P_{40}	P_{50}	P_{60}	P_{70}	P_{75}	P_{80}	P_{90}
中位數						M_e					
四分位數			Q_1			Q_2			Q_3		
十分位數	D_1	D_2		D_3	D_4	D_5	D_6	D_7		D_8	D_9

根據上述的百分位數之觀念，我們亦可推出其他**分位數** (fractiles) 的觀念。例如，第一個三分位數即約略等於 $P_{33.33}$，而第二個三分位數約略等於 $P_{66.67}$。就例 3.8 來看，其位置指標 $i = \frac{66.67}{100} \times 50 = 33.335$，不為整數，故取第 34 個位置上的數值 63.1，亦即第二個三分位數等於 63.1。

著作權所有 侵害必究
華泰文化 HWA TAI PUBLISHING

3.2 差異量數

前一節所述的集中趨勢量數基本上是在衡量一組資料之集中趨勢或中心位置，而**差異量數**(dispersion measures)則在衡量一組資料中，各個觀測值之間的差異或離散的程度（故差異量數亦稱為離散量數），由此可表現出此組資料的散布範圍及分布的形態，並可反映出平均數的代表性。舉個例子來說，設有下列三組學生的成績：

甲組	80	80	80	80	80
乙組	76	78	80	82	84
丙組	50	70	90	90	100

此三組學生的平均成績雖皆為 80 分，但各組成績之差異與分布情形則顯然不同。甲組五人的成績每人均為 80 分，毫無差異；乙組中五人之成績雖有差別，但差別尚不甚大；而丙組五人的成績中有特別高或特別低者，顯示學生程度彼此間差異很大。由此可知，平均數相同的資料，其差異情形未必相同。因此，為了解一群數值之較完整的特性，除了平均數外，尚須測定其差異量數。

本節將介紹幾個比較重要的差異量數，包括全距、四分位差、平均偏差、標準差（與變異數）。這些差異量數基本上都有其優缺點，使用時應視資料特性以及應用目的而選擇較合適的量數。

著作權所有 侵害必究

3.2.1 全距

一組資料中，數值最大者與最小者之差稱為**全距**(range)，一般以英文字母 R 表示。以全距來測度資料的差異程度有二個優點，即意義簡明易解以及計算容易。然而，其嚴重的缺點為，若資料所包含的觀測值非常大或非常小時，其敏感度相當小；而且全距無法測出中間各個觀測值之間的差異情形。

例 3.9

設有二組資料如下：

A： 3, 4, 5, 6, 7, 9, 9, 10, 12, 15
B： 3, 8, 8, 9, 9, 9, 10, 15

試求出其全距、平均數與中位數，並做比較。

華泰文化 HWA TAI PUBLISHING