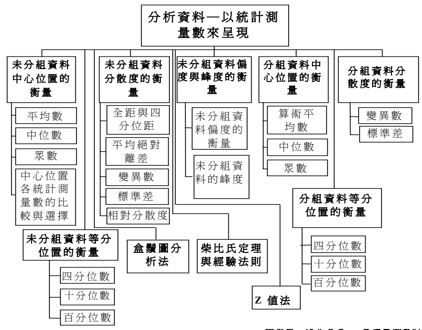
5　分析資料–以統計測量數來呈現

學習目的

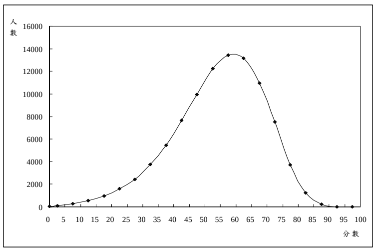
1. 瞭解資料中心位置的各種衡量指標如算術平均數、中位數、眾數、加權平均數與幾何平均數等的衡量方法。
2. 熟習各個中心位置衡量方法的特性、使用時機與優缺點。
3. 瞭解資料分散程度的各種衡量指標如全距、四分位距、變異數、標準差、變異係數的衡量方法。
4. 熟習各分散程度衡量方法的特性、使用時機與優缺點。
5. 認識資料等分位置的的各種衡量方法如四分位數、十分位數百分位數等的計算。認識與計算資料的偏度、峰度。
6. 熟習使用EXCEL計算中心位置與分散度指標及等分位置之指標。

本章結構



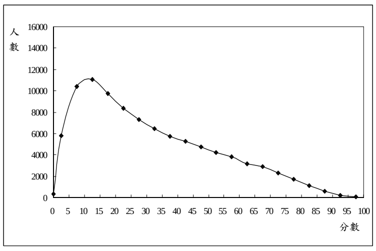
分析資料–以統計測量數來呈現

圖5.1　九十六學年度指定學科考試國文科成績



分析資料–以統計測量數來呈現

圖5.2　九十六學年度指定學科考試英文科成績



分析資料–以統計測量數來呈現

表5.1　九十六學年度指定學科考試國文與英文成績的統計測量數

	國文科	英文科
平均數	54.4422	31.0917
變異數	211.4321	464.9560
標準差	14.5407	21.5628
眾數	57.5000	12.5000
中位數	53.4756	23.7921

未分組資料中心位置的衡量

- 平均數

算術平均數的意義

所有觀察值的總和除以觀察值的個數即爲算術平均數。算術平均數在數線上代表資料的平衡點。

母體平均數

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

樣本平均數

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

未分組資料中心位置的衡量

表5.2　銀行業與證券業的月薪

	A	B	C	D	E	F	G	H
1	證券業	20	23	23	25	26	29	64
2	銀行業	26	27	27	28	30	32	33

未分組資料中心位置的衡量

- **算數平均數的特質**
  - ①資料的平衡點
  - ②各觀察值與平均數間的差的總和最小
  - ③各觀察值與平均數之差的平方和最小
  - ④優點爲考慮到每一個觀察值，缺點爲易受極端值的影響。
  - ⑤可進行代數演算
  - ⑥可對觀察值予以加權

未分組資料中心位置的衡量

圖5.3 證券業的平均月薪

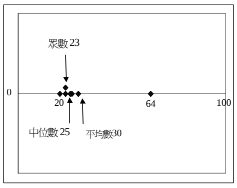
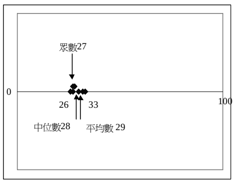


圖5.4 銀行業的平均月薪



未分組資料中心位置的衡量

- **平均數**

  加權算術平均數

  母體：$\mu_w = \dfrac{\sum_{i=1}^{N} W_i x_i}{\sum_{i=1}^{N} W_i}$　　樣本：$\bar{X}_w = \dfrac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i}$

未分組資料中心位置的衡量

表5.3　學生成績報告單

	A	B	C	D	E
1	科目	學分數	學期成績	加權成績	
2	國文	3	86	258	
3	英文	3	87	261	
4	歷史	2	95	190	
5	體育	1	87	87	
6	微積分	3	82	246	
7	經濟學原理	4	86	344	
8	會計學	3	89	267	
9	社會心理學	3	90	270	
10	工程發展與社會變遷	2	91	182	
11		24		2105	87.71

未分組資料中心位置的衡量

- **樣本的幾何平均數**

$$\bar{g} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

未分組資料中心位置的衡量

表5.4 台積電的股票價格

年度	台積電股價	變動比
90	77.74	
91	67.42	0.867
92	56.42	0.837
93	52.36	0.928
94	54.08	1.033
95	61.34	1.134
96	65.52	1.068

未分組資料中心位置的衡量

❍ **幾何平均數的性質**

① $\sqrt[n]{\prod_{i=1}^{n}(x_i / y_i)} = \sqrt[n]{\prod_{i=1}^{n} x_i} \Big/ \sqrt[n]{\prod_{i=1}^{n} y_i}$

例如 X 為國民所得，Y 為人口數，則 X/Y 為平均每人所得，要求算其平均成長率可以計算 X/Y 的幾何平均數，或分別計算 X 與 Y 的幾何平均數，再將兩個幾何平均數相除。

②適合衡量等比數列的中央位置，但不易進行統計推論。

未分組資料中心位置的衡量

❍ **幾何平均數的投資報酬率**

$$\overline{G} = \left[(1+R_1)(1+R_2)\cdots(1+R_n)\right]^{\frac{1}{n}} - 1$$

未分組資料中心位置的衡量

❍ **中位數**

中位數是位於依數值大小順序排列的觀察值中央的那一個數值。

未分組資料中心位置的衡量

❍ **中位數的特質**

① $\sum|x_i - m_e|$ 為 $\sum|x_i - A|$ 中之最小，亦即 $\sum|x_i - m_e| \le \sum|x_i - A|$（A 為任意數）。此乃意指一組觀察值中，若欲尋找一個代表值使觀察值與代表值的距離和為最小，則該代表值即為中位數。

②不受極端值的影響。中位數只是觀察值數列中的一個數值，因此當然不受極端值的影響，故對觀察值的變化不敏感。

③不易進行代數演算，亦不易進行統計推論。

未分組資料中心位置的衡量

❍ **眾數**

眾數是指觀察值中其出現次數最多的那一個數值或類別。

未分組資料中心位置的衡量

❍ 眾數的性質

①不受極端值的影響

②可能有多個或一個也沒有

③對觀察值的個數或數值變化的感應不靈敏

④眾數因可能有多個或一個也沒有，因此眾數比中位數及平均數較少使用。

未分組資料中心位置的衡量

表4.4　中央趨勢統計測量數之比較

統計測量數	優　點	缺　點
算術平均數	1.資料的重心。資料無極端值或偏態時，具代表性。 2.適合代數演算 3.考慮所有觀察值，敏感度高。 4.觀察值與平均數差平方和最小 5.適合統計推論的工作	1.若有極端值存在時則不具代表性 2.資料如爲偏態，則代表性較差。
幾何平均數	1.適合等比資料 2.敏感度高	1.不適合一般資料 2.不適合作統計推論
中位數	1.適用於有極端值的資料 2.適用於偏態資料 3.觀察值與中位數絕對差和最小 4.可做無母數統計推論	1.不適合代數演算 2.對觀察值敏感性低 3.不易進行母數統計推論
眾數	1.適用於有極端值的資料 2.適用於偏態資料 3.適用於質的資料	1.可能不止一個或不存在 2.敏感性低 3.不能做統計推論

未分組資料中心位置的衡量

表5.6　士林與桃園地院訴訟案件審理日數

士林地院	桃園地院
65　110　58　72　66　67　68　65	73　78　64　77　63　73
78　90　82　86	105　61　81　59

未分組資料中心位置的衡量

表5.7　地院訴訟案件審理日數的平均數中位數與眾數

	A	B	C	D	E
1	士林地院		桃園地院		
2					
3	平均數	75.58	平均數	73.4	
4	中間值	70	中間值	73	中位數
5	眾數	65	眾數	73	

未分組資料等分位置(分割數）的衡量

❍ **四分位數**

四分位數是將順序資料分成四等分數值的分位數。四分爲數有第1、第2、第3三個寺分位數。

❍ **十分位數**

十分位數是將資料均分爲十等份數值的分割數。

❍ **百分位數**

百分位數是將順序資料均分爲一百等分數值的分割數。

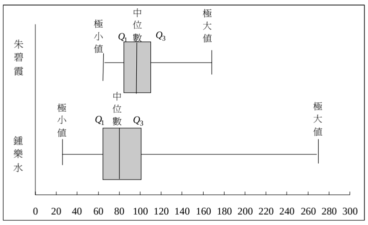
未分組資料等分位置(分割數）的衡量

產業經濟學的學期成績

78	79	80	81	82	83	83	84	84
85	86	87	88	89	90	91	92	95

盒鬚圖分析法(5數綜合)

圖5.5 業務員業績的盒鬚圖



未分組資料分散度的衡量

- **全距**

  $R = \text{最大值} - \text{最小值}$

- **四分位距**

  $IQR = \text{第3四分位數} - \text{第1四分位數} = Q_3 - Q_1$

- **平均絕對離差**

  母體：$MAD = \frac{1}{N}\sum_{i=1}^{N} | x_i - \mu |$

  樣本：$mad = \frac{1}{n}\sum_{i=1}^{n} | x_i - \overline{X} |$

未分組資料等分位置(分割數)的衡量

圖5.6 A股票價格的分配

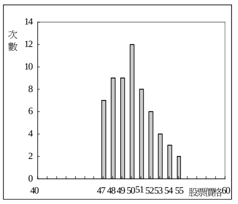
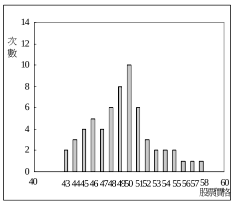
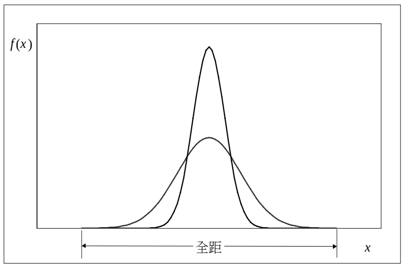


圖5.7 B股睞價格的分配



未分組資料分散度的衡量

圖5.8 全距相同但分散程度不同



未分組資料分散度的衡量

表5.9 縱貫路與中山高的開車時間

	A	B	C	D	E	F
1	縱貫路	37	34	39	38	42
2	中山高	44	23	37	31	55

未分組資料分散度的衡量

表5.10 縱貫路開車時間的平均絕對離差

開車時間	$X - \overline{X}$	$\lvert X - \overline{X} \rvert$	$(X - \overline{X})^2$
37	−1	1	1
34	−4	4	16
39	1	1	1
38	0	0	0
42	4	4	16
合計	0	10	34

未分組資料分散度的衡量

- **變異數**

母體變異數

$$\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$$

式中：μ：母體平均數，N：母體個數。

樣本變異數

$$S^2 = \frac{1}{n-1}\sum(x_i - \overline{X})^2$$

式中：$\overline{X}$：樣本平均數，n：樣本數。

未分組資料分散度的衡量

- **變異數的性質**

①變異數的值大於等於0，若變異數為0時，其意義是所有觀察值均相同，沒有變異（分散）。
②若同一組資料單位不同，其變異數亦不相同。
③單位相同可作比較
④考慮每一個觀察數值
⑤適合代數演算
⑥適合利用樣本變異數對母體變異數做統計推論
⑦具有複名數（如元^2），不易解釋。如電腦價格的變異數的單位為平方元（元^2），不具意義。

未分組資料分散度的衡量

- **標準差**

母體標準差

$$\sigma = \sqrt{\sigma^2}$$

樣本標準差

$$S = \sqrt{S^2}$$

未分組資料分散度的衡量

表5.11　縱貫路與中山高開車時間的比較

	A	B	C	D
1	縱貫路開車時間		中山高開車時間	
2				
3	平均數	38	平均數	38
4	中間值	38	中間值	37
5	眾數	#N/A	眾數	#N/A
6	標準差	2.92	標準差	12.25
7	變異數	8.5	變異數	150

未分組資料分散度的衡量

- **相對分散度**

變異係數

$$\text{變異係數}(CV) = \frac{\text{標準差}}{\text{平均數}}$$

母體資料：$CV = \dfrac{\sigma}{\mu}$

樣本資料：$CV = \dfrac{S}{\overline{X}}$

未分組資料中央趨勢的衡量

表5.12　兩種基金報酬率的平均數與標準差

基金類別	平均數（%）	標準差（%）
甲基金	11.32	6.63
乙基金	7.21	4.87

柴比氏定理與經驗法則

❍柴比氏定理

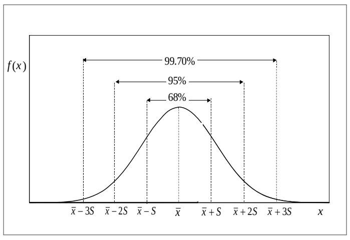
不論資料為何種分配，至少有$(1-1/k^2)$的資料落在距離平均數 k 個標準差的範圍內。k 為大於1的任意數，即$k>1$。

❍經驗法則

若資料為鐘形分配，則有68%的觀察值落在$\bar{x}\pm S$內，有95%的觀察值落在$\bar{x}\pm 2S$內，有99.7%的觀察值落在$\bar{x}\pm 3S$內（S為標準差）。

柴比氏定理與經驗法則

圖5.9　經驗法則



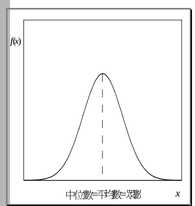
Z值

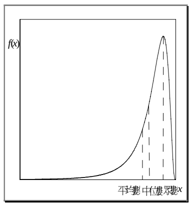
❍ Z 值

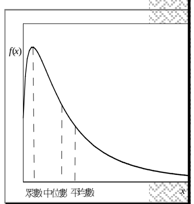
樣本x值的 Z 值：$\dfrac{x-\bar{x}}{s}$　，母體 X 值的 Z 值：$\dfrac{X-\mu}{\sigma}$

未分組資料偏度與峰度的衡量

圖5.10 對稱分配　**圖5.11 左偏分配**　**圖5.12 右偏分配**







未分組資料偏度與峰度的衡量

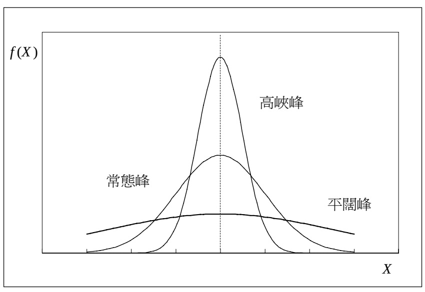
❍ 皮爾生偏態係數

母體：$SK_p=\dfrac{3(\mu-M_e)}{\sigma}$

樣本：$SK_p=\dfrac{3(\bar{X}-m_e)}{S}$

未分組資料偏度與峰度的衡量

圖5.13 三種峰度的圖形



未分組資料偏度與峰度的衡量

- **如何檢查極端值（outliers）**

  所謂極端值是指與其他大部分的數值比較起來爲極小或極大的數值，利用下列步驟可檢查是否有極端值。

  步驟 1：將觀察值由小而大排列

  步驟 2：計算出第一四分位數Q_1與第三四分位數Q_3

  步驟 3：計算四分位距 $IQR = Q_3 - Q_1$

  步驟 4：計算 $Q_1 - 1.5 \times IQR$ 及 $Q_3 + 1.5 \times IQR$

  步驟 5：若觀察值 x 小於 $Q_1 - 1.5 \times IQR$ 或大於 $Q_3 + 1.5 \times IQR$ 則爲極端值。

分組資料中心位置的衡量

- **算術平均數**

  母體均數：$\mu = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$　　樣本均數：$\overline{X} = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

分組資料中心位置的衡量

表5.13　海之濱營業收入的次數分配表

組號	組限	組距	組中點	次數 f_i
1	$20 \le x < 30$	10	25	4
2	$30 \le x < 40$	10	35	7
3	$40 \le x < 50$	10	45	12
4	$50 \le x < 60$	10	55	18
5	$60 \le x < 70$	10	65	11
6	$70 \le x < 80$	10	75	6
7	$80 \le x < 90$	10	85	3
				$\sum f_i = 61$

分組資料中心位置的衡量

表5.14　海之濱營業收入的次數分配表

	A	B	C	D	E	F	G
1	組號	組限	組距	組中點 x_i	次數 f_i	$f_i x_i$	累加次數
2	1	$20 \le x < 30$	10	25	4	100	4
3	2	$30 \le x < 40$	10	35	7	245	11
4	3	$40 \le x < 50$	10	45	12	540	23
5	4	$50 \le x < 60$	10	55	18	990	41
6	5	$60 \le x < 70$	10	65	11	715	52
7	6	$70 \le x < 80$	10	75	6	450	58
8	7	$80 \le x < 90$	10	85	3	255	61
9					61	3,295	

分組資料中心位置的衡量

- **中位數**

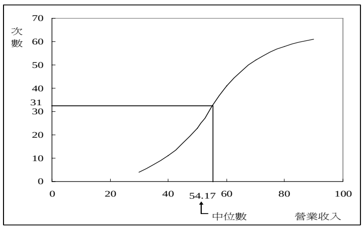
  $$m_e = L_{m_e} + W_{m_e}\left(\frac{\frac{n}{2} - F_L}{f_{m_e}}\right)$$

  式中：L_{m_e}：m_e 所在組的組下界，W_{m_e}：m_e 所在組的組距，

  f_{m_e}：m_e 所在組的組次數，F_L：m_e 前一組的累加次數。

分組資料中心位置的衡量

圖5.14　海之濱每日平均營業收入中位數的圖解



分組資料中心位置的衡量

- 眾數

  粗略法眾數

$$m_0 = \frac{(\text{組上界} + \text{組下界})}{2}$$

分組資料等分位置的衡量

- 四分位數

  第1四分位數

$$Q_1 = L_{Q_1} + \frac{\frac{1}{4}n - F_{Q_1}}{f_{Q_1}} W_{Q_1}$$

式中：L_{Q_1}：Q_1所在組的組下界，f_{Q_1}：Q_1所在組的組次數，
W_{Q_1}：Q_1所在組的組距，F_{Q_1}：Q_1前一組的累加次數。

第3四分位數

$$Q_3 = L_{Q_3} + \frac{\frac{3}{4}n - F_{Q_3}}{f_{Q_3}} W_{Q_3}$$

式中：L_{Q_3}：Q_3所在組的組下界，f_{Q_3}：Q_3所在組的組次數，
W：Q_3所在組的組距，F_{Q_3}：Q_3前一組的累加次數。

分組資料等分位置的衡量

表5.15 學生英文考試成績次數分配表

組號	組限	次數	累加次數
1	30~40	2	2
2	40~50	1	3
3	50~60	12	15
4	60~70	14	29
5	70~80	38	67
6	80~90	33	100
7	90~100	6	106

分組資料等分位置的衡量

- 十分位數

$$D_i = L_{D_i} + \frac{\frac{n \cdot i}{10} - F_{D_i}}{f_{D_i}} W_{D_i}$$

式中：D_i：第i個十分位數，L_{D_i}：D_i所在組的組下界，f_{D_i}：D_i所在組的組次數，W_{D_i}：D_i所在組的組距，F_{D_i}：D_i前一組的累加次數。

- 百分位數

$$P_i = L_{P_i} + \frac{\frac{n \cdot i}{100} - F_{P_i}}{f_{P_i}} W_{P_i}$$

式中：P_i：第i個百分位數，L_{P_i}：P_i所在組的組下界，f_{P_i}：P_i所在組的組次數，W_{P_i}：P_i所在組的組距，F_{P_i}：P_i前一組的累加次數。

分組資料分散度的衡量

- 變異數與標準差

  母體變異數與標準差

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 f_i \qquad \sigma = \sqrt{\sigma^2}$$

式中：x_i：組中點，f_i：組次數，N：母體個數，k：組數。

樣本變異數與標準差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{X})^2 f_i \qquad S = \sqrt{S^2}$$

式中：x_i：組中點，f_i：組次數，n：母體個數，k：組數。

分組資料分散度的衡量

表5.16 海之濱營業收入的變異數與標準差

	A	B	C	D	E	F
1	組號	組限	組中點 x_i	次數 f_i	$x-\overline{X}$	$(x-\overline{X})^2 f_i$
2	1	$20 \le x < 30$	25	4	-29.02	3,368.64
3	2	$30 \le x < 40$	35	7	-19.02	2,532.32
4	3	$40 \le x < 50$	45	12	-9.02	976.32
5	4	$50 \le x < 60$	55	18	0.98	17.29
6	5	$60 \le x < 70$	65	11	10.98	1,326.16
7	6	$70 \le x < 80$	75	6	20.98	2,640.96
8	7	$80 \le x < 90$	85	3	30.98	2,879.28
9				$\sum = 61$		13,740.98

EXCEL 的使用

衡量資料的中心位置的測量數如算術平均數、中位數、眾數，衡量分散程度的測量數如標準差、變異數等，除可利用Excel的「資料分析」中的「敘述統計」來進行外，尚可利用Excel的「插入函數」功能（開啟Excel，然後選取「公式」、「插入函數」，再選取類別（統計或數學與三角函數等）、接著「選取函數）來進行。「敘述統計」雖比較方便，一次可獲得許多摘要性統計數據，但是它對某些統計測量數卻沒有提供。因此像絕對離差（AVEDEV）、四分位數（QUARTILE）、百分位數（PERCENTILE）等的衡量，只能利用「插入函數」功能來求算。

第5章 分析資料–以統計測量數來呈現 應用統計學四版

EXCEL 的使用

表5.17 海之濱春季的營業收入

	A	B	C
1	海之濱春季的營業收入		
2			
3	平均數	54	
4	中間值	54	中位數
5	眾數	45	
6	標準差	15.344	
7	變異數	235.433	
8	峰度	-0.301	
9	偏態	0.068	
10	範圍	68.000	全距
11	最小值	21.000	
12	最大值	89.000	
13	個數	61.000	元素個數

林惠玲 陳正倉著 雙葉書廊發行 2009