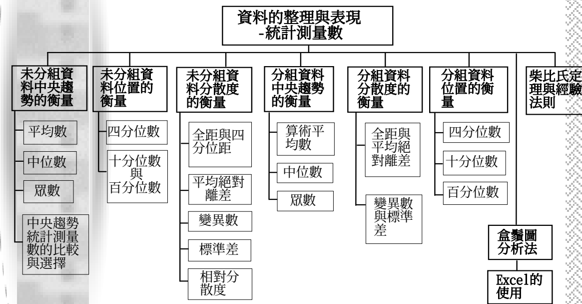
4 資料的整理與表現-統計測量數

學習目的

1. 了解資料中央趨勢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. 熟習各個中央趨勢衡量方法的特性、使用時機與優缺點。
3. 了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. 熟習各個分散程度衡量方法的特性、使用時機與優缺點。
5. 認識資料相對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計算中央趨勢與分散度指標及其他位置之指標。

本章結構



未分組資料中央趨勢的衡量

❍ 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為算術平均數。算術平均數在數線上代表資料的平衡點。

母體平均數

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

樣本平均數

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

圖4.1 板橋的平均業績

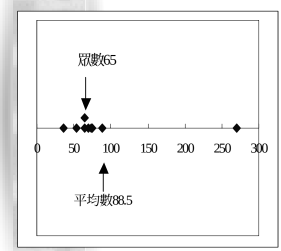
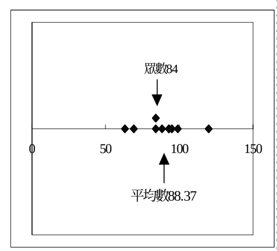


圖4.1 北投的平均業績



未分組資料中央趨勢的衡量

❍ 平均數

加權算術平均數

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

母體的幾何平均數

$$G = \sqrt[N]{x_1 x_2 \cdots x_N} = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

樣本的幾何平均數

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

算術平均數的特質

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優點為考慮到每一個觀察值，缺點為易受極端值的影響。

⑤可進行代數演算

⑥可對觀察值予以加權

表4.1 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	81	243	
3	英文	3	82	246	
4	哲學概論	3	88	264	
5	中國通史	2	90	180	
6	西洋哲學史	2	89	178	
7	中國憲法與政治	2	86	172	
8	經濟學原理	3	78	234	
9	體育	1	75	75	
10	法學緒論	2	82	164	
11				1756	83.61905

表4.2 消費者物價指數表（幾何平均數）

年別	消費者物價指數	變動率
85	100.00	
86	100.90	1.0090
87	102.60	1.0168
88	102.78	1.0018
89	104.07	1.0126

資料來源：Taiwan Statistical Data Book 2001，行政院經建會。

消費者物價指數變動率為 $CPI_t\ /\ CPI_{t-1}$

未分組資料中央趨勢的衡量

- 中位數

  中位數的意義

  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。

- 眾數

  眾數的意義

  眾數是指觀察值中其出現次數最多的那一個數值。

圖4.3 板橋的平均業績

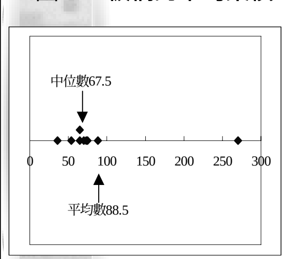


圖4.4 北投的平均業績

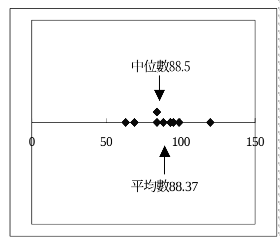


表4.3 中央趨勢統計測量數之比較

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.資料如為偏態，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

表4.4 平均數中位數與眾數

	A	B	C	D	E	F
1	板橋業績		北投業績			
2						
3	平均數	88.5	平均數	88.375		
4	標準誤	26.30793	標準誤	6.318729	平均數的標準差	
5	中間值	67.5	中間值	88.5	中位數	
6	眾數	65	眾數	84		

未分組資料位置的衡量(其他測量數)

- **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。

- **十分位數**

  十分位數是將資料均分為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為一百等分數值的分割數。

圖4.6 四分位數的圖解

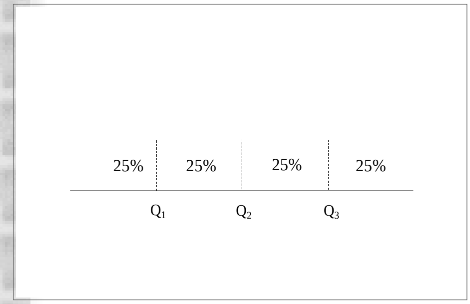


圖4.6 甲廠商交貨期的分配

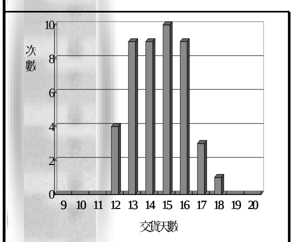
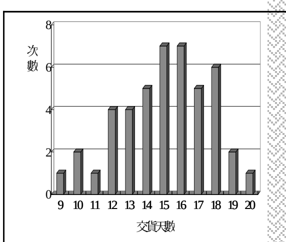


圖4.7 乙廠商交貨期的分配



未分組資料分散度的衡量

- **全距**

  $R=$ 最大值 $-$ 最小值

- **四分位距**

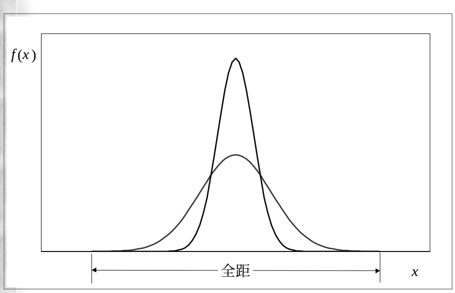
  $IQR=$ 第3四分位數 $-$ 第1四分位數 $=Q_3-Q_1$

- **平均絕對離差**

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

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

圖4.8 資料的分散情形



縱貫路	40	42	43	44	46
中山高	21	36	42	55	61

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

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

式中：$\bar{X}$ ：樣本平均數，n ：樣本數。

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

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

表4.6 縱貫路與中山高開車時間的比較

	A	B	C	D
1	縱貫公路		中山高	
2				
3	平均數	43	平均數	43
4	標準誤	1	標準誤	7.078135
5	中間值	43	中間值	42
6	眾數	#N/A	眾數	#N/A
7	標準差	2.236068	標準差	15.82719
8	變異數	5	變異數	250.5

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：$CV = \dfrac{S}{\bar{X}}$

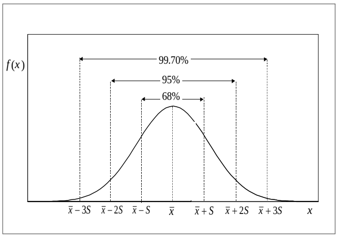
表4.7 台積電與聯電股票的投資報酬率

	A	B	C
1	台積電	聯電	
2	11.2	8.2	
3	12.2	-6.3	
4	-1.7	20.4	
5	7.6	-2.8	
6	19.9	35	
7	25.4	44.7	
8	26.3	25.9	
9	14.2	6.3	
10	-2.7	2.8	
11	12.4	30.5	

圖4.8 兩種股票的投資報酬率

	A	B	C	D
1	台積電		聯電	
2				
3	平均數	12.48	平均數	16.47
4	標準誤	3.114439	標準誤	5.467074
5	中間值	12.3	中間值	14.3
6	眾數	#N/A	眾數	#N/A
7	標準差	9.848722	標準差	17.28841
8	變異數	96.99733	變異數	298.889

圖4.13 經驗法則



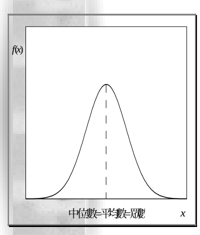
未分組資料的衡量-偏態與峰度

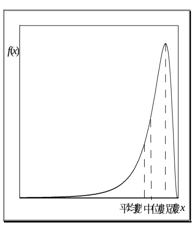
- 皮爾生偏態係數

母體：$SK_p = \dfrac{3(\mu - M_e)}{\sigma}$

樣本：$SK_P = \dfrac{3(\bar{X} - m_e)}{S}$

圖4.9 對稱分配 圖4.10 左偏分配 圖4.11 右偏分配





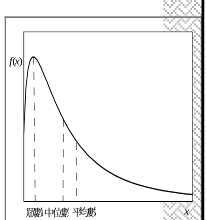
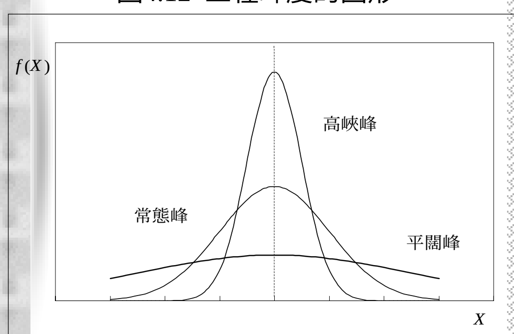


圖4.12 三種峰度的圖形



分組資料中央趨勢的衡量

- 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：$\bar{X} = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

- 中位數

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

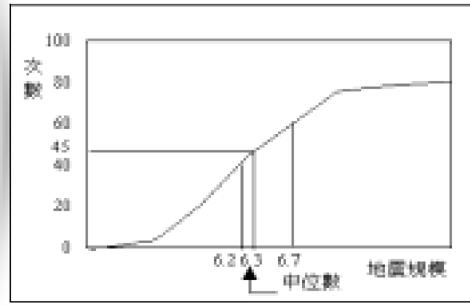
式中：L_{m_e}：m_e 所在組的組下界，W_{m_e}：m_e 所在組的組距，

f_{m_e}：m_e 所在組的組次數，F_L：m_e 前一組的累加次數。

表4.11 20世紀台灣災害性地震規模的次數分配表

組別	組限	組界	組距	組中點	次數 f_i
1	$4.5 \le x \le 4.9$	$4.45 \le x < 4.95$	0.5	4.7	1
2	$5.0 \le x \le 5.4$	$4.95 \le x < 5.45$	0.5	5.2	4
3	$5.5 \le x \le 5.9$	$5.45 \le x < 5.95$	0.5	5.7	19
4	$6.0 \le x \le 6.4$	$5.95 \le x < 6.45$	0.5	6.2	26
5	$6.5 \le x \le 6.9$	$6.45 \le x < 6.95$	0.5	6.7	16
6	$7.0 \le x \le 7.4$	$6.95 \le x < 7.45$	0.5	7.2	18
7	$7.5 \le x \le 7.9$	$7.45 \le x < 7.95$	0.5	7.7	3
8	$8.0 \le x \le 8.4$	$7.95 \le x < 8.45$	0.5	8.2	2
					$\sum f_i = 89$

圖4.13 台灣地區災害性地震規模中位數的圖解



分組資料中央趨勢的衡量

- 眾數

粗略法眾數

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

分組資料分散度的衡量

- 全距

① $R =$ 最後一組組中點 $-$ 第一組組中點

② $R =$ 最後一組組上限 $-$ 第一組組下限

- 平均絕對離差

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

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

x_i：組中點，k為組數，f_i：組次數。

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。

表4.13 地震規模的變異數與標準差

	A	B	C	D	E	F	G
1	組別	組限	x_i組中點	f_i次數	$x-\overline{X}$	$\lvert x-\overline{X}\rvert f_i$	$(x-\overline{X})^2 f_i$
2	1	4.5≤x≤4.9	4.7	1	-1.7	1.7	2.89
3	2	5.0≤x≤5.4	5.2	4	-1.2	4.8	5.76
4	3	5.5≤x≤5.9	5.7	19	-0.7	13.3	9.31
5	4	6.0≤x≤6.4	6.2	26	-0.2	5.2	1.04
6	5	6.5≤x≤6.9	6.7	16	0.3	4.8	1.44
7	6	7.0≤x≤7.4	7.2	18	0.8	14.4	11.52
8	7	7.5≤x≤7.9	7.7	3	1.3	3.9	5.07
9	8	8.0≤x≤8.9	8.2	2	1.8	3.6	6.48
10				Σ=89		51.7	43.51

分組資料位置的衡量（其他測量數）

- 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。

表4.14 上市公司股價的次數分配表

組號	組限	組距	次數	累加次數
1	$0 \le x < 2$	2	30	30
2	$2 \le x < 4$	2	54	84
3	$4 \le x < 6$	2	61	145
4	$6 \le x < 8$	2	56	201
5	$8 \le x < 10$	2	47	248
6	$10 \le x < 15$	5	78	326
7	$15 \le x < 25$	10	89	415
8	$25 \le x < 40$	15	57	472
9	$40 \le x < 60$	20	32	504
10	$60 \le x < 100$	40	13	517
11	$100 \le x < 250$	150	11	528

資料來源：《臺灣證券交易所上市證券概況》，90年6月。

分組資料位置的衡量（其他測量數）

- **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。

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%的觀察值落在$\bar{X} \pm 3S$內（S為標準差）。

表4.11 經驗法則

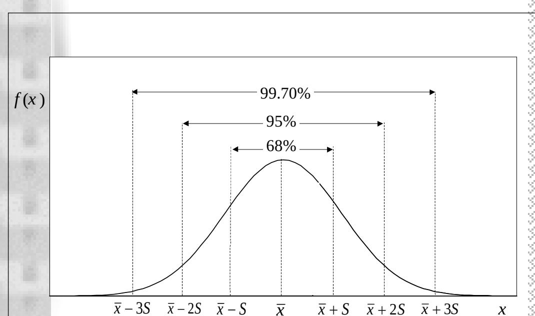
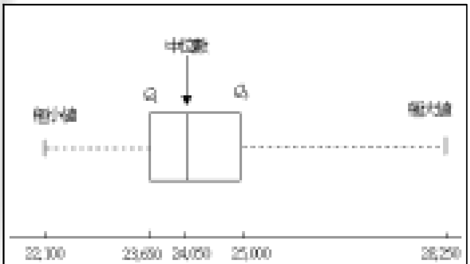


圖4.15 盒鬚圖



兩變數間的關係

- 共變數

母體共變數

$$COV(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\mu_X)(y-\mu_Y)}{N}$$

樣本共變數

$$cov(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n-1}$$

圖4.19 正向共變

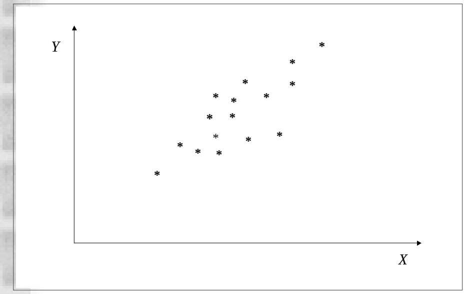


圖4.20 反向共變

兩變數間的關係

- 相關係數

母體相關係數

$$\rho_{XY}=\frac{COV(X,Y)}{\sigma_X\sigma_Y}$$

式中：$COV(X,Y)$ 為母體共變數, σ_X , σ_Y 為標準差。

樣本相關係數

$$r_{XY}=\frac{cov(X,Y)}{S_X S_Y}$$

式中 $cov(X,Y)$ 為樣本共變數, S_X , S_Y 為樣本標準差。

偏度

- 皮爾生偏度係數

母體：$SK_p=\frac{3(\mu-M_e)}{\sigma}$

樣本：$SK_P=\frac{3(\overline{X}-m_e)}{S}$

圖4.25 三種峰度的圖形