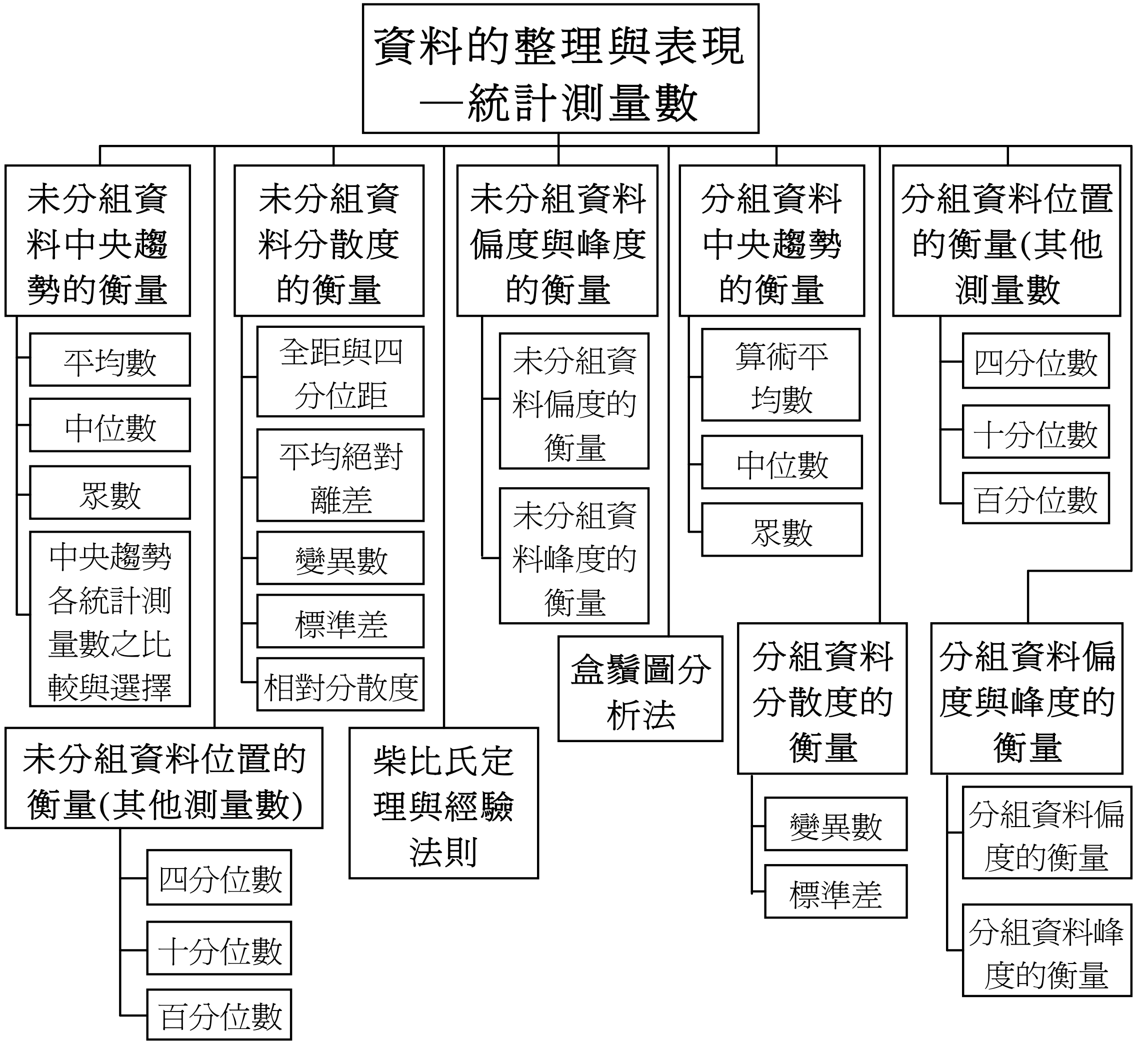
第4章　資料的整理與表現-統計測量數

林惠玲　陳正倉著 雙葉書廊發行 2004

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



林惠玲 陳正倉著 雙葉書廊發行 2004

4.1 未分組資料中央趨勢的衡量

目的: 用來描述資料的中心位置或中央趨勢

測量數: 平均數 (mean), 中位數 (median), 眾數 (mode)

4.1.1 平均數

代表一組資料的平均水準

指標:

- 算術平均數 (arithmetic mean)
- 幾何平均數 (geometric mean)
- 調和平均數 (harmonic mean)

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

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

樣本平均數

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

林惠玲 陳正倉著 雙葉書廊發行 2004

例4.1 北投好還是板橋強 (以平均數做代表)

某汽車公司現有北投與板橋二個營業處, 2003年8月板橋營業處與北投營業處業務員的業績依序為(單位: 萬元)

試問: 板橋及北投的業務員其平均業績各為何?
板橋營業處的業績好還是北投營業處的業績強

板橋營業處	32	52	64	64	70	76	82	280
北投營業處	61	72	81	81	95	97	101	124

林惠玲 陳正倉著 雙葉書廊發行 2004

板橋

$$\mu = \frac{32 + 52 + \ldots + 280}{8} = 90$$

北投

$$\mu = \frac{61 + 72 + \ldots + 124}{8} = 89$$

從平均業績來看, 板橋與北投的業績差不多

◆算術平均數的特質 (六點)

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

1. 算術平均數是資料的平衡點

所謂的資料坪衡點，是指平均數左邊的觀察值與平均數的距離的總和，等於平均數右邊的觀察值與平均數的距離的總和

圖4.1 板橋的平均業績

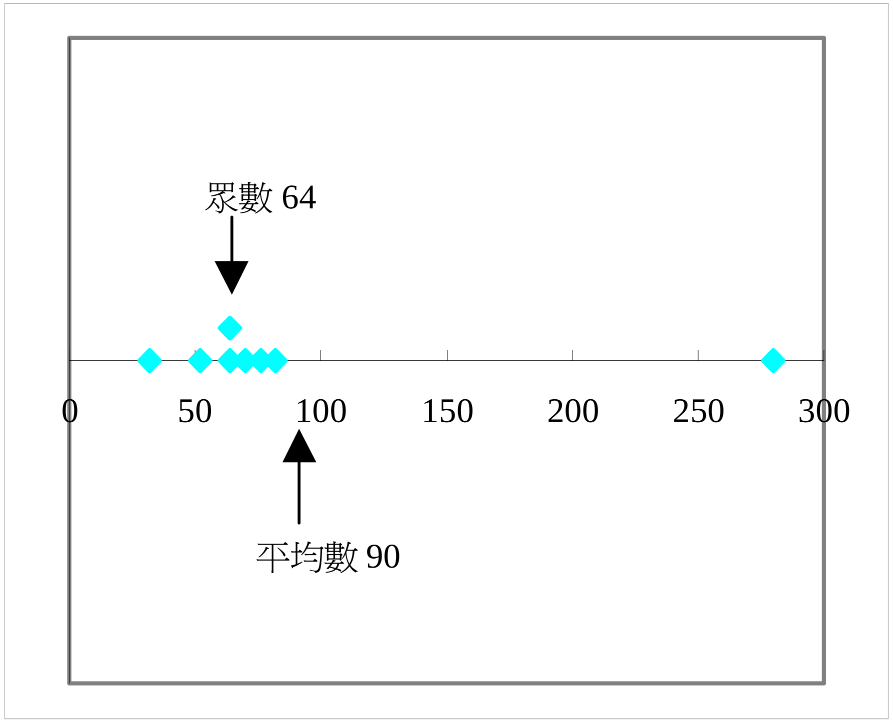
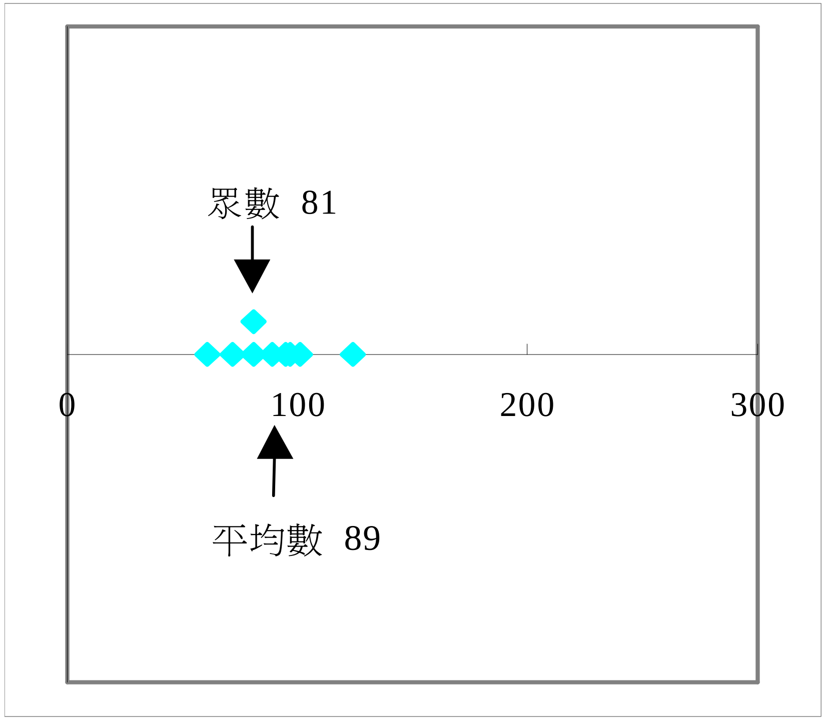


圖4.2 北投的平均業績



2. 各觀察值與平均數間的差的總和等於零

$$\sum_{i=1}^{N}(x_i-\mu)=0 \quad \text{or} \quad \sum_{i=1}^{N}(x_i-\overline{X})=0$$

3. 各觀察值與平均數之差的平方和最小

$$\sum_{i=1}^{N}(x_i-\mu)^2 \quad 在 \quad \sum_{i=1}^{N}(x_i-\mathrm{A})^2=0 \quad 爲最小$$

$$\text{or} \quad \sum_{i=1}^{N}(x_i-\overline{\mathrm{X}})^2 \quad 在 \quad \sum_{i=1}^{N}(x_i-\mathrm{A})^2=0 \quad 爲最小$$

在一群觀察值中,若欲尋找一個代表值,使觀察值與代表值誤差平方和為最小,則該代表值即為平均數

4. 優點是使用到每一個觀察值,缺點是易受極端值影響

ex: 板橋營業處拿掉超級營業員(280), 平均數由90降為62.86

5. 可進行代數的演算

若X的線性函數為 Y=a+bX

則 $\mu_Y = a + b\mu_X$ $\qquad \overline{Y} = a + b\overline{X}$

Ex: 設有兩組資料, X與Y. 已知X之母體平均數為μ_X, Y之母體平均數為 μ_Y, 則X與Y總平均μ為

$$\mu = \frac{N_X \mu_X + N_Y \mu_Y}{N_X + N_Y}$$

林惠玲 陳正倉著 雙葉書廊發行 2004

6. 可對觀察值予以加權 (加權平均數)

概念: 算術平均數視每一個觀察值為一樣重要

→ 當觀察值有不同的重要性時

→ 對每一個觀察值給予一個權數(weight)，用以代表其重要性, 然後再計算其平均數

→ 加權的算術平均數 (weighted arithmetic mean)

母體

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

樣本

$$\bar{X}_W = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i}$$

Ex4.4 某學生大一上學期的成績如表**4.1,** 求其學其平均成績

	A	B	C	D	E
1	科目	學分數	成績	加權成績	平均成績
2	國文	3	85	255	
3	英文	3	93	279	
4	歷史	2	76	152	
5	體育	1	87	87	
6	微積分	3	91	273	
7	經濟學原理	4	83	332	
8	會計學	3	92	276	
9	藝術欣賞	2	86	172	
10	合計	21		1826	86.95

林惠玲 陳正倉著 雙葉書廊發行 2004

幾何平均數

◇ 若資料為等比數列, 如國民生產毛額成長率, 物價上漲率…, 則應以幾何平均數來代表該等資料的中心位置

母體的幾何平均數

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

樣本的幾何平均數

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

表4.2　台灣塑膠公司的股票價格

年度	台塑	變動率
87	46.5	
88	62.5	1.344
89	46.3	0.741
90	32.1	0.693
91	44.6	1.421
92	56	1.228

資料來源：台灣証券交易所。註：價格的變動率為 P_t / P_{t-1}

試求台灣塑膠公司股價變動的幾何平均數及算術平均數，再求幾何平均數報酬率

Solutions

台塑股價變動率的幾何平均為

$$\bar{g}_{台塑}=(1.344\times0.741\times0.693\times1.421\times1.228)^{\frac{1}{5}}=1.038$$

台塑股票的幾何投資報酬率為

$$R_{台塑}=1.038-1=0.038$$

台塑股票的算術平均數的投資報酬率為

$$\bar{X}_{台塑}=[(1.344+0.741+0.693+1.421+1.228)/5]-1=0.085$$

以算術平均數來看為高估

幾何平均數的性質

1.

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

2. 適合衡量等比數列的中央位置,但不易進行統計推論

幾何平均數的應用

幾何平均數投資報酬率

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

式中 Ri 為第i期的投資報酬率

Measure of Central Tendency For The Rate Of Change Of A Variable Over Time: The Geometric Mean & The Geometric Rate of Return

- Geometric mean
  - Used to measure the rate of change of a variable over time

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return
  - Measures the status of an investment over time

$$\bar{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

  - Where R_i is the rate of return in time period i

The Geometric Mean Rate of Return: Example

An investment of $100,000 declined to $50,000 at the end of year one and rebounded to $100,000 at end of year two:

$$\underbrace{X_1 = \$100,000 \quad X_2}_{\text{50\% decrease}} = \underbrace{\$50,000 \quad X_3 = \$100,000}_{\text{100\% increase}}$$

The overall two-year return is zero, since it started and ended at the same level.

The Geometric Mean Rate of Return: Example

(continued)

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\overline{X} = \frac{(-.5) + (1)}{2} = .25 = 25\%$$

Misleading result

Geometric mean rate of return:

$$\begin{aligned}\overline{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1 \\ &= [(1+(-.5))\times(1+(1))]^{1/2} - 1 \\ &= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More representative result

修正的平均數

1. 剪尾平均數

將兩尾α%的觀察值去掉, 只利用100%-2α%的觀察值求取平均數

2. 截尾平均數

以P_{α}(第α的百分位數)值代替P_{α}以下的觀察值, 且以$P_{100-\alpha}$值代替$P_{100-\alpha}$右尾α的觀察值, 再求其平均數

林惠玲 陳正倉著 雙葉書廊發行 2004

4.1.2 中位數

☆定義:

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。

☆算法:

設X為變數, 其母體觀察值由小到大排列
為 $x_1 < x_2 < \cdots < x_N$,N為觀察值個數

a) 若N為奇數, 則中位數m_e為數列中第 (N+1)/2個的那一個數值

b) 若N為偶數, 則取數列中第 (N+1)/2前後兩個數之平均數為中位數

☆性質

1. $\sum|x_i - m_e|$ 爲 $\sum|x_i - A|$ 中之最小, 亦即 $\sum|x_i - m_e| \le \sum|x_i - A|$ (A爲任意數值). 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最小,則該代表值即爲中位數

2. 不受極端值影響 (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

林惠玲 陳正倉著 雙葉書廊發行 2004

例4.7

Q: 在例4.1中已知板橋及北投業務員的業績, 現問板橋及北投業務員業績的中位數各爲何?

Solution:

n=8, 中位數應位於第(n+1)/2 = (8+1)/2 =4.5個 前後各一個之平均數

故取第4及第5個觀察值的平均數

板橋

$$m_e = (64+70)/2 = 67$$

北投

$$m_e = (81+95)/2 = 88$$

北投因無極端值存在, 平均數89與中位數88較爲接近

圖4.3 板橋的平均業績

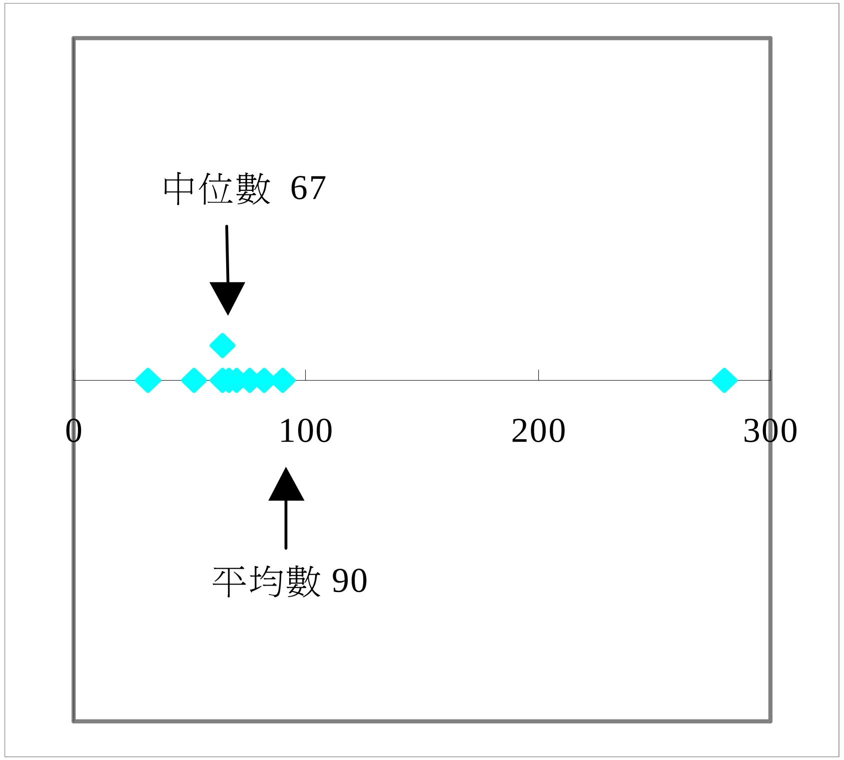
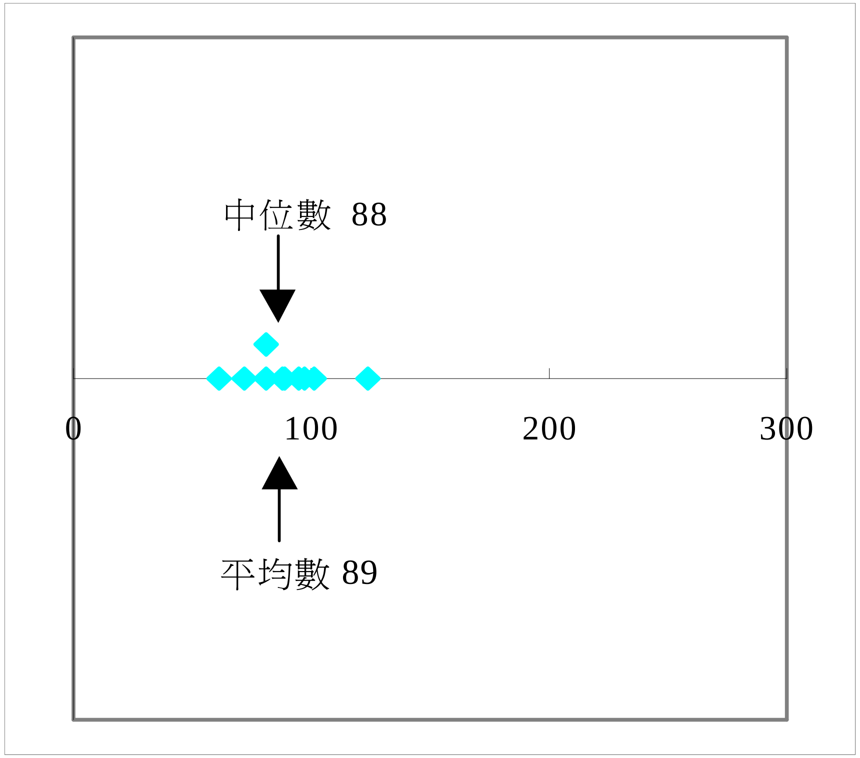


圖4.4 北投的平均業績



4.1.3 眾數

☆定義:

眾數是觀察值中出現次數最多的那一個數值獲類別。

以m_o表示

☆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沒有, 因此眾數比中位數及平均數較少使用

表4.3 肺癌患者的存活時間

患者人數	存活天數
3	3天
3	1年
2	3年
1	10年
1	20年

資料來源：虛擬。

例4.9

看法	維持現狀以後再決定	永遠維持現狀	維持現狀以後統一	維持現狀以後獨立	儘快宣布獨立	儘快統一
人數	358	231	111	108	63	45
百分比	39.05%	24.22%	12.12%	11.79%	6.88%	4.91

資料來源：行政院陸委會91年12月委託e社會資訊管理公司調查。

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

林惠玲　陳正倉著 雙葉書廊發行 2004

表4.5 板橋與北投業績的平均數、中位數與眾數

	A	B	C	D	E
1	板橋業績		北投業績		
2					
3	平均數	90	平均數	89	
4	中間值	67	中間值	88	中位數
5	眾數	64	眾數	81	

1. 開啟Excel工作表,將資料輸入
2. 選取 “工具”,“資料分析”,“敘述統計” 即可得

4.2 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。

4.2.1 四分位數

☆定義:

四分位數是將順序資料的觀察值分成四等分數值的分位數, 亦即, 四分位數有第1,第2, 第3三個四分位數

☆求算:

a) 第1四分位數(Q_1)

先計算K=n/4, 若K爲整數, 則取第K個與第K+1個兩個數值的平均數爲第1四分位數. 若K不是整數, 則取小數進位爲整數的那一個

b) 第2四分位數 (Q_2). 第2四分位數就是中位數

c)第3四分位數(Q_3)

先計算K=3n/4, 若K爲整數, 則取第K個與第K+1個兩個數值的平均數爲第3四分位數. 若K不是整數, 則取小數進位爲整數的那一個

4.2.2 十分位數

☆定義:

十分位數是將觀察值均分為十等分數值的分割數,

亦即, 十分位數有9個, 第i個表示為D_i(i=1,…,9)

☆求算:

先將觀察值由小而大排列, 並求算$K=\dfrac{n\cdot i}{10}$, i=1,…9, n為觀察值的個數, 此後再依下面兩個情況來決定十分位數

a)若K為整數, , 則取第K個與第K+1個兩個數值的平均數為第 i 個十分位數D_i

b)若K不是整數, 則取小數進位為整數的那一個數值為第 i 個十位數D_i

例4.12 下面為台灣本國壽險公司2000年的保費收入(已順序排列),試求第5個和第7個十分位數的保費收入為多少?

1,847	2,665	2,835	4,539	4,865	6,168	11,763
14,930	18,730	18,851	31,594	100,821	107,610	204,840

Solution

$$K_5 = \frac{n \cdot i}{10} = \frac{14 \times 5}{10} = 7 \qquad K_7 = \frac{n \cdot i}{10} = \frac{14 \times 7}{10} = 9.8$$

故第5個十分位數為 (11763+14930)/2=13346.5
故第7個十分位數為 18851百萬元

林惠玲 陳正倉著 雙葉書廊發行 2004

4.2.3 百分位數

☆定義:

百分位數是將順序資料的觀察值均分爲一佰等份數值的分割數,亦即, 十分位數有99個, 第i個表示爲P_i(i=1,…,99)

☆求算:

百分位數的求算方法與十分位數相同,不同的是百分位數

公式爲 $P=\frac{n\cdot i}{100}$ i=1,…,99. 而十分位數爲 $K=\frac{n\cdot i}{10}$ i=1,..,9

4.3 未分組資料分散度的衡量

概念

平均數,中位數與眾數僅表示資料的中心位置(集中或聚集情形), 未能呈現資料的分散情形.但有時分散程度的重要性不亞於中央趨勢.

EX: 一個中心廠有底下二個配合的衛星工廠, 選哪一個較佳?

→ 雖然二工廠的交貨期均為15天,但乙工廠的交貨期較為分散,對中心廠而言,生產流程的安排會造成某些問題

☆測量數

全距 (Range, R), 四分位距, 平均絕對離差(average absolute deviation), 變異數 (variance), 標準差

變異係數(coefficient of variance, CV, 衡量相對分散度)

林惠玲 陳正倉著 雙葉書廊發行 2004

圖4.5 甲廠商交貨期的分配

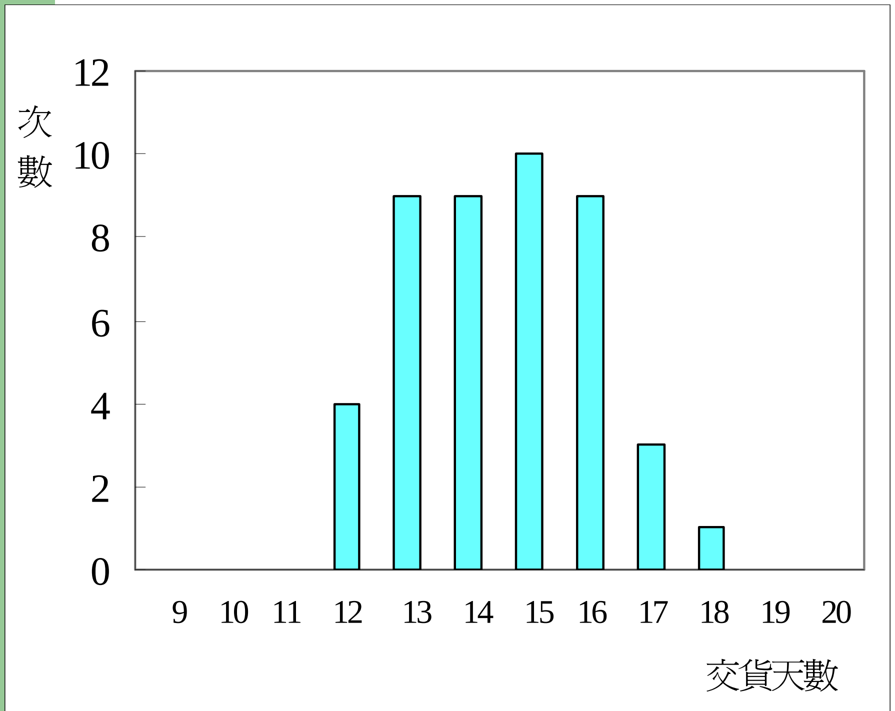
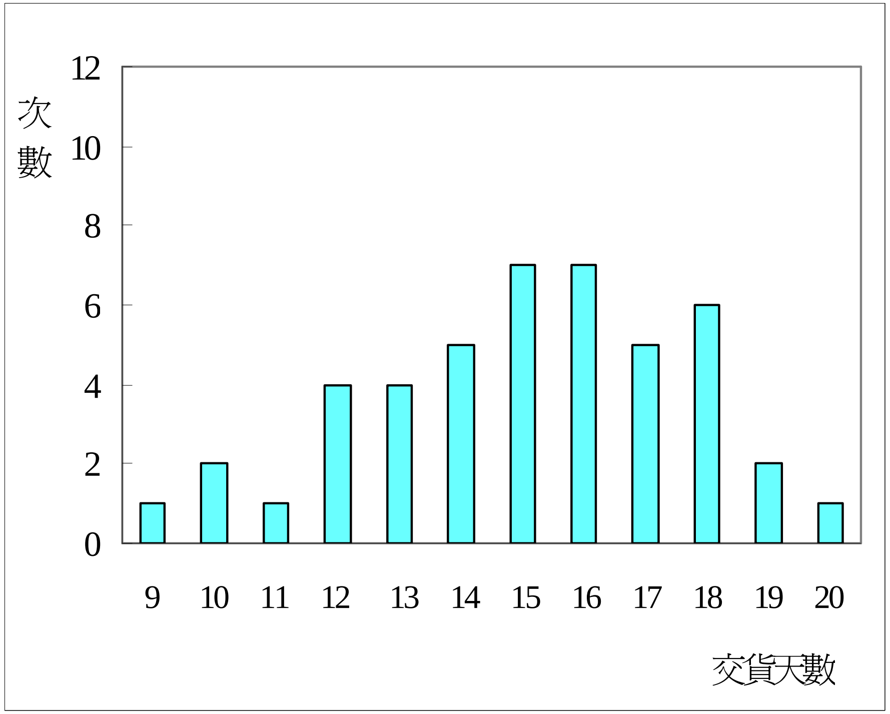


圖4.6 乙廠商交貨期的分配



4.3.1 全距與四分位距

全距 (Range, R)

觀察值中的最大值減去最小值後的數值

R=最大值 – 最小值 (R越大, 表分散程度越大)

四分位距

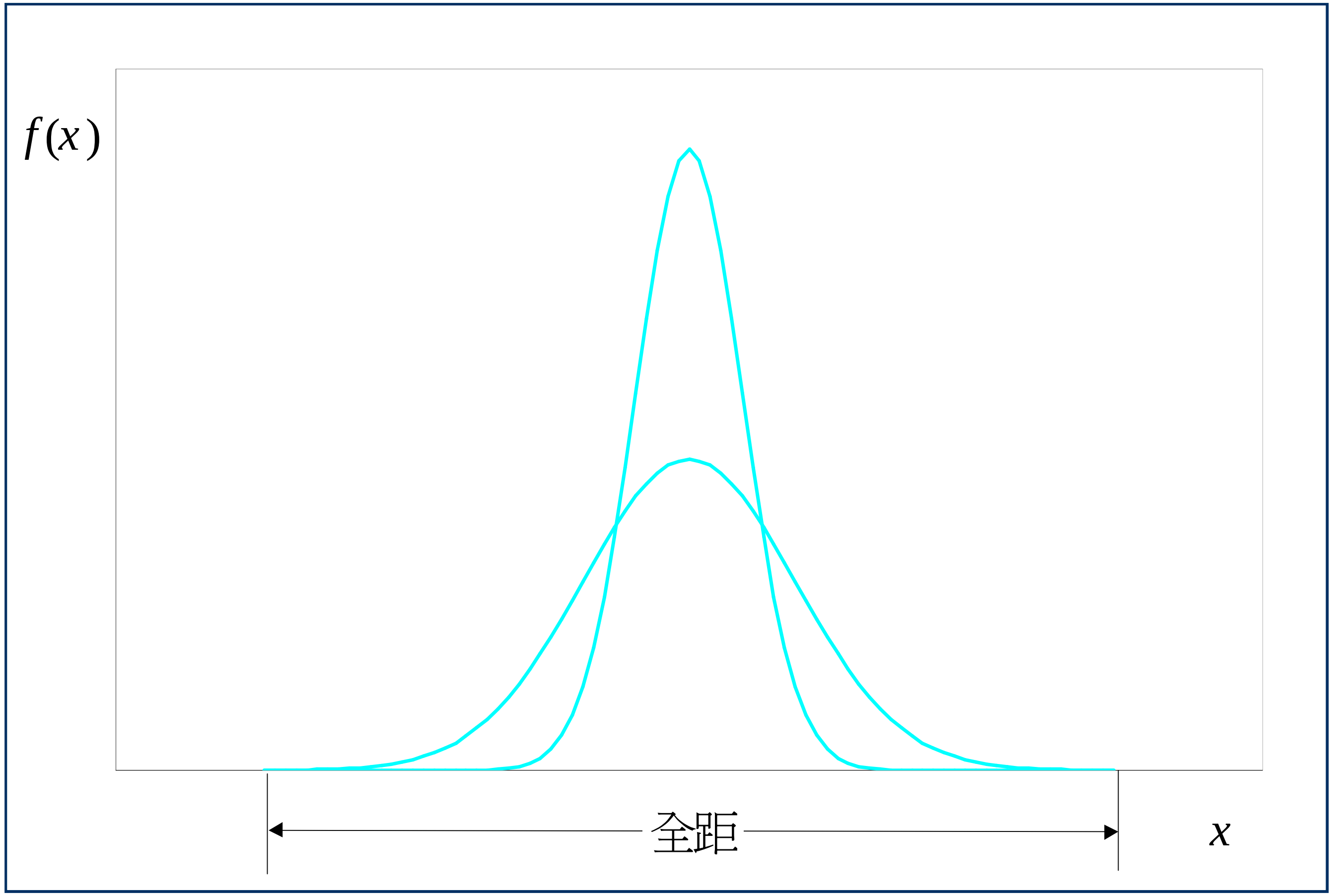
第3四分位數減去第1四分位,

$$IQR = Q_3 – Q_1$$

☆ 全距的缺點 (四分位距亦同)

1. 資料的單位不同,不能比較
2. 當資料單位相同,且全距相同時,我們也不能下結論說兩個資料的分散程度相同 (如下圖4.7)
3. 全距只考慮最大值與最小兩個觀察值,未考慮所有觀察值, 故不能精確的反應全體觀察值的分散情形,會受極端值影響

圖4.7　全距相同時資料的分散情形



4.3.2 平均絕對離差

定義: 每一個觀察值與平均數之間的差距, 稱為平均數的離差 (deviation about the mean), $x_1 - \overline{X}, x_2 - \overline{X}, \ldots, x_n - \overline{X}$

平均絕對離差

母體: $MAD = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \mu\right|$ 樣本: $mad = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \overline{X}\right|$

☆性質

平均絕對離差值越大表示分散程度越高

例4.14 走哪一條路好

北部地區15歲以上通勤通學民眾平日通勤時間平均為1.15小時. 洪小姐由新莊到台北市上班有二條路: 一是經三重到台北, 一是經泰山走中山高到台北,她紀錄2個星期的行車時間如下(分鐘), 問洪小姐走哪條路較好?

縱貫路	40	42	43	44	46
中山高	21	36	42	55	61

Solution:

可得走二條路的平均時間均為43分鐘
省道之全距為46-40=6 平均絕對離差為1.6分鐘
中山高全距為61-21=40, 平均絕對離差為12分鐘

→省道差異程度小

林惠玲 陳正倉著 雙葉書廊發行 2004

表4.6 縱貫路開車時間的平均絕對離差

開車時間	$X-\bar{X}$	$\lvert X-\bar{X}\rvert$	$(X-\bar{X})^2$
40	−3	3	9
42	−1	1	1
43	0	0	0
44	1	1	1
46	3	3	9
合計	0	8	20

4.3.3 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。

☆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. 單位相同可做比較
4. 考慮每一個觀察數值
5. 適合代數運算 (見下頁)
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

代數運算

a) Y=a+bX, X的母體變異數爲 σ_X^2 ,標準差爲σ_X

則Y的母體變異數與標準差分別爲 σ_Y^2 與σ_Y

$$\sigma_Y^2 = b^2\sigma_X^2,\ \sigma_Y = |b|\sigma_X$$

b) 兩組資料X與Y, X母體變異數與標準差分別爲μ_X與 σ_X^2

Y母體變異數與標準差分別爲μ_Y與 σ_Y^2

則兩組資料之全體變異數爲

$$\sigma^2 = \frac{N_X\left[\sigma_X^2 + (\mu_X - \mu)^2\right] + N_Y\left[\sigma_Y^2 + (\mu_Y - \mu)^2\right]}{N_X + N_Y}$$

其中

$$\mu = \frac{N_X\mu_X + N_Y\mu_Y}{N_X + N_Y}$$

N爲個數

4.3.4 標準差

概念: 變異數具複名數,不易解釋,未去除此缺點而將變異數開根號所得的數值稱爲標準差

母體標準差

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

樣本標準差

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

林惠玲 陳正倉著 雙葉書廊發行 2004

例4.15 省道與中山高開車時間的變異數與標準差

依上述公式

省道變異數

$$S^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2 = \frac{20}{5-1} = 5$$

省道標準差

$$S = \sqrt{S^2} = \sqrt{5} = 2.236$$

中山高變異數

$$S^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2 = \frac{1002}{5-1} = 250.5$$

中山高標準差

$$S = \sqrt{S^2} = \sqrt{250.5} = 15.827$$

洪小姐選哪一條? 必須看他的時間和個性來決定. 若他個性溫和不急躁,且爲了避免遲到的麻煩,她應選省道並早點出門

4.3.5 相對分散度

概念: 若有兩組資料而欲比較其相對分散程度時,如何去除單位的不同來進行比較??

變異係數

$$變異係數\ (CV) = \frac{標準差}{平均數}$$

$$母體資料:\ CV = \frac{\sigma}{\mu}$$

$$樣本資料:\ CV = \frac{S}{\overline{X}}$$

林惠玲　陳正倉著 雙葉書廊發行 2004

表4.7　兩種基金的平均數與標準差

請分析下面兩種股票型基金的投資風險, 並問哪種基金的投資風險較大?

基金類別	平均數（%）	標準差（%）	基金個數
跨國投資全球型	10.28	6.03	20
開放式價值類	5.09	3.71	5

資料來源：台灣經濟新報。

林惠玲　陳正倉著 雙葉書廊發行 2004

Solution:

跨國投資全球型 平均報酬率10.28%, 標準差6.03%

開放式價值類基金 平均報酬率5.06%, 標準差3.71%

跨國投資全球型 CV=6.03/10.28 = 0.59

開放式價值類基金 CV=3.71/5.09 = 0.73

比較兩個變異數可知, 低投資報酬率基金(開放式價值類基金)的變異細數大於高投資報酬率基金(跨國投資全球型)的變異係數,這個結果與高報酬高風險的現象不一致

4.4 柴比氏定理與經驗法則

❍ 柴比氏定理(Chebyshev's Theorem)

不論資料爲何種分配，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爲大於1的任意數，即$k>1$。

❍ 經驗法則(empirical rule)

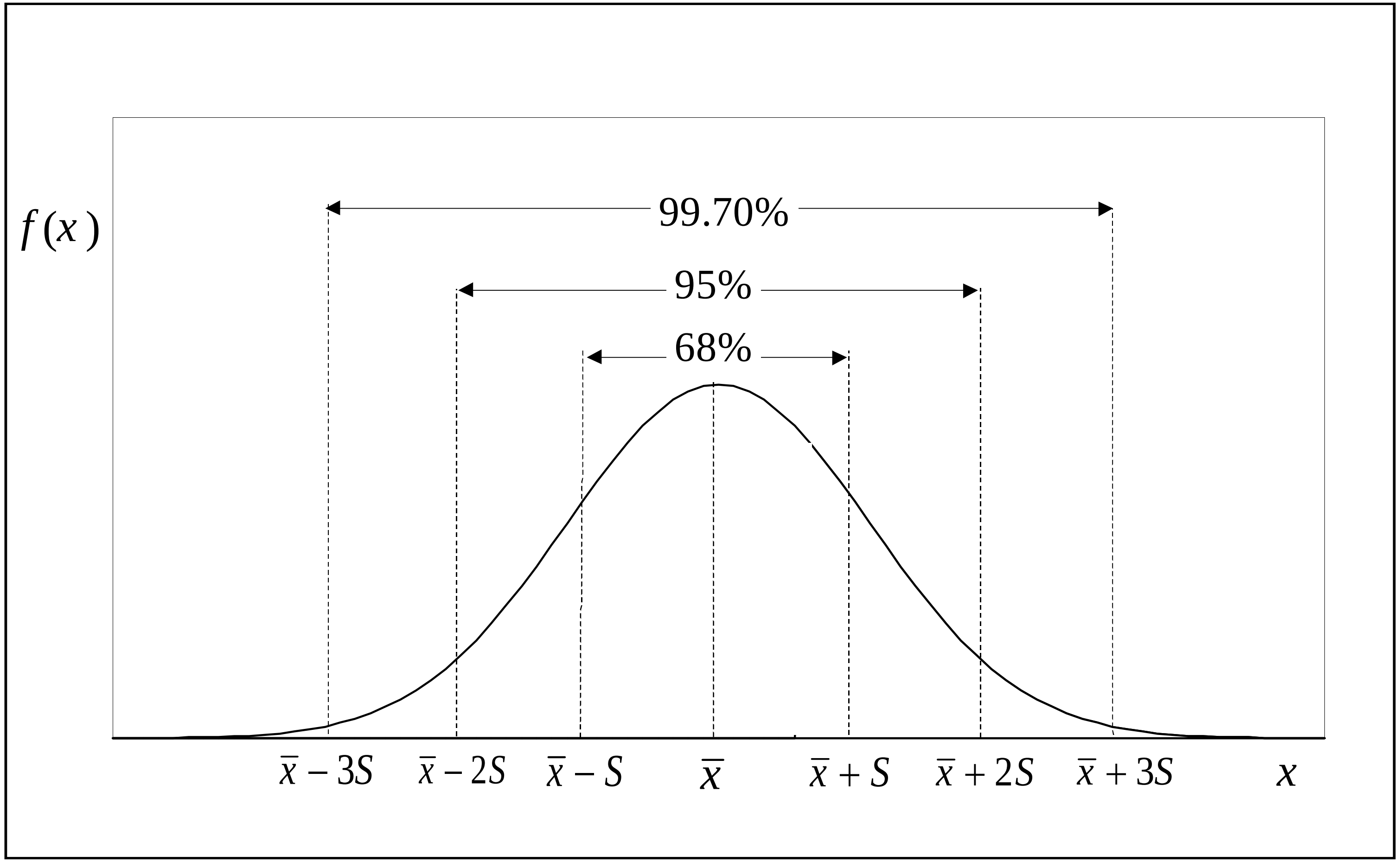
若資料爲鐘形分配(bell-shaped distribution)，則有68%的觀察值落在$\bar{X}\pm S$內，有95%的觀察值落在$\bar{X}\pm 2S$內，有99%的觀察值落在$\bar{X}\pm 3S$內（S爲標準差）。

林惠玲 陳正倉著 雙葉書廊發行 2004

How the standard deviation provides a measure of variability of a data set? How many observations fit within $\underline{+}$ n s of the mean?

	Chebyshev's Rule	Empirical Rule
$\pm 1s$ or $\pm 1\sigma$	No useful info	Approximately 68%
$\pm 2s$ or $\pm 2\sigma$	At least 75%	Approximately 95%
$\pm 3s$ or $\pm 3\sigma$	At least 8/9	Approximately 99.7%

圖4.8 經驗法則



4.5 未分組資料偏度與峰度的衡量

概念: 資料的分布可能爲不對稱 (左偏或右偏)

4.5.1 未分組資料偏態的衡量

偏態的方向可分爲對稱,左偏或右偏三種

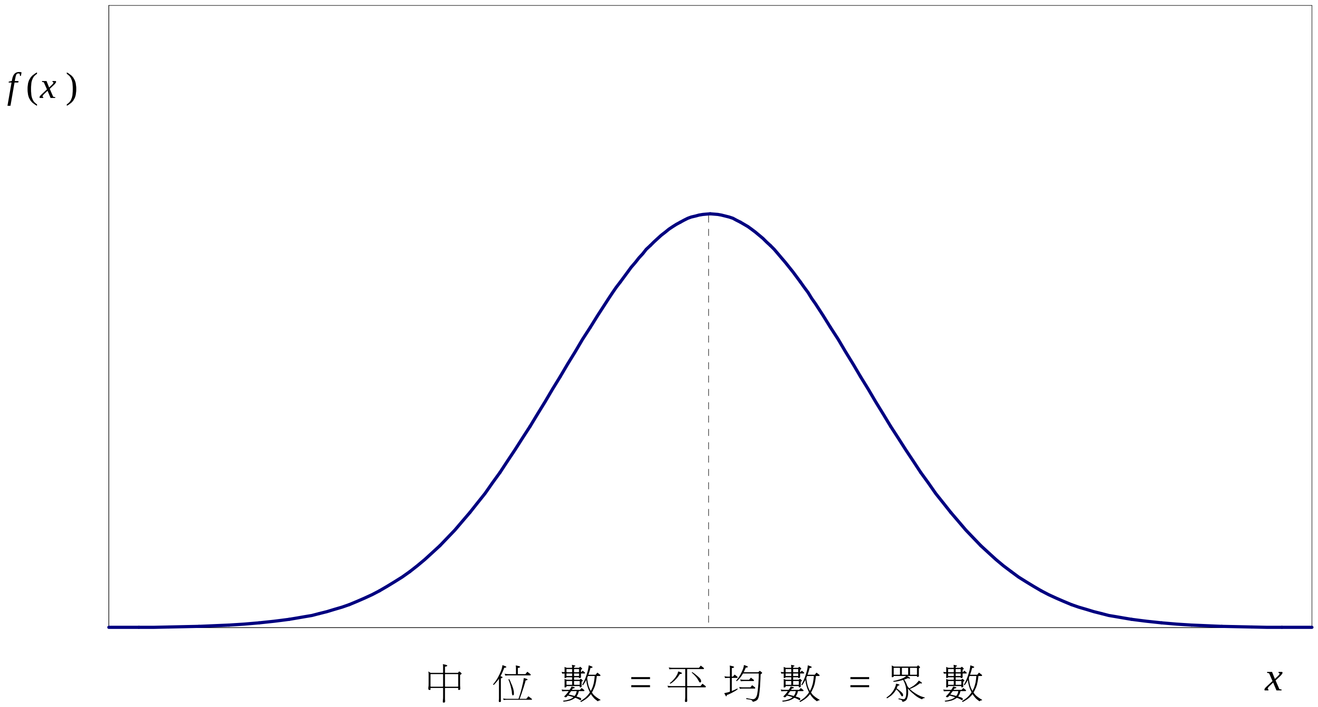


圖4.10 左偏分配

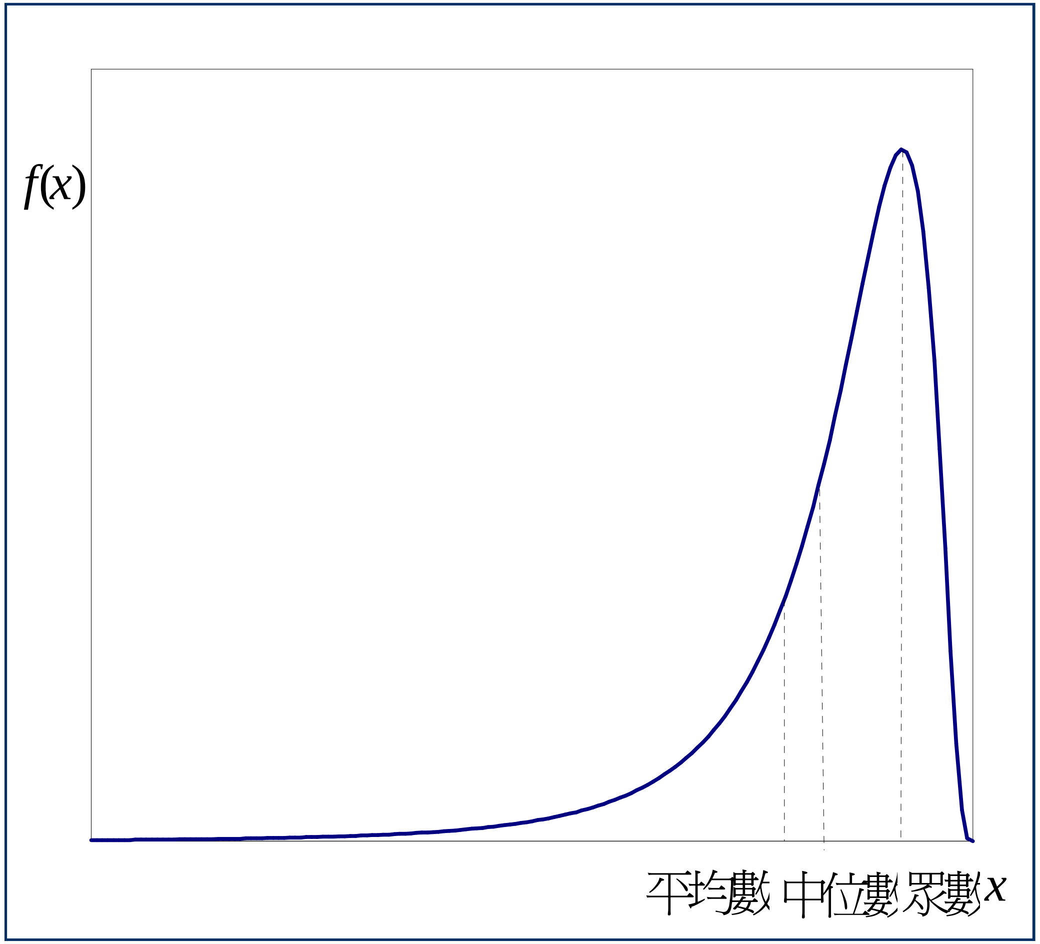
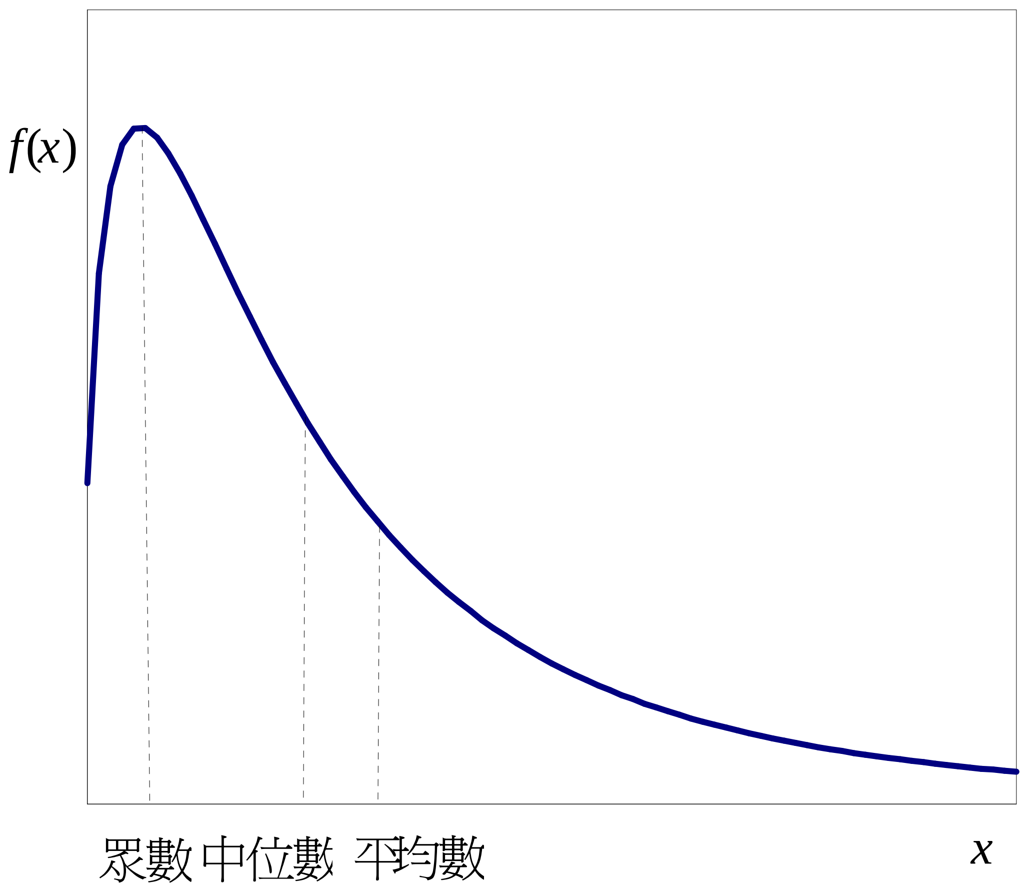


圖4.11 右偏分配



如何衡量偏態?

皮爾生偏態係數 (Pearson skewness coefficient)

母體：

$$SK_p = \frac{\mu - M_o}{\sigma} = \frac{3(\mu - M_e)}{\sigma}$$

樣本：

$$SK_p = \frac{\bar{X} - m_o}{S} = \frac{3(\bar{X} - m_e)}{S}$$

$SK_p=0$ 對稱分配 $\bar{X} = m_o$

$SK_p>0$ 右偏分配 $\bar{X} > m_o$

$SK_p<0$ 左偏分配 $\bar{X} < m_o$

例4.17 板橋營業處的銷售額是否爲偏態

Solution:

已知 $\bar{X} = 90, m_o = 64, S = 78$

$$SK_P = \frac{3(\bar{X} - m_o)}{S} = \frac{90-64}{78} = 0.33$$

由此知,板橋營業處的銷售額爲略微右偏

☆ 動差法的偏態係數

動差法的母體偏態係數

$$\alpha_3 = \frac{M_3}{\sigma^3} = \frac{M_3}{\left(\sqrt{M_2}\right)^3}$$

動差法的樣本偏態係數

$$\alpha_3 = \frac{m_3}{(\sqrt{m_2})^3}$$

☆ 性質

1. $\alpha_3=0$ 為對稱分配 $\alpha_3>0$ 為右偏分配, $\alpha_3<0$ 為左偏分配
2. $|\alpha_3|$ 越大表示越偏態

$0 \le |\alpha_3| \le 0.5$ 趨於對稱分配 $0.5 < |\alpha_3| < 1$ 稍微偏態

$|\alpha_3| > 1$ 極為偏態

母體平均數的動差

$$M_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^r$$

1. 一級動差 r=1, 則 $M_1 = \frac{1}{N}\sum(x-\mu) = 0$ 母體平均離均差
2. 二級動差 r=2, 則 $M_2 = \frac{1}{N}\sum(x-\mu)^2 = \sigma^2$ 母體變異數
3. 三級動差 r=3, 則 $M_3 = \frac{1}{N}\sum(x-\mu)^3$ 衡量偏態
4. 四級動差 r=4 則 $M_4 = \frac{1}{N}\sum(x-\mu)^4$ 衡量峰態

例 4.18 信用卡刷卡金額是否為偏態

10位中大經濟系學生的台新銀行信用卡刷卡金額如下(元)

4,798 2,192 2,085 1,211 3,570 1,548 875 250 158

Solution:

樣本平均數 $\bar{X} = 1{,}781.7$

$$m_2 = \frac{1}{N}\sum(x_i - \bar{X})^2 = 2{,}034{,}441$$

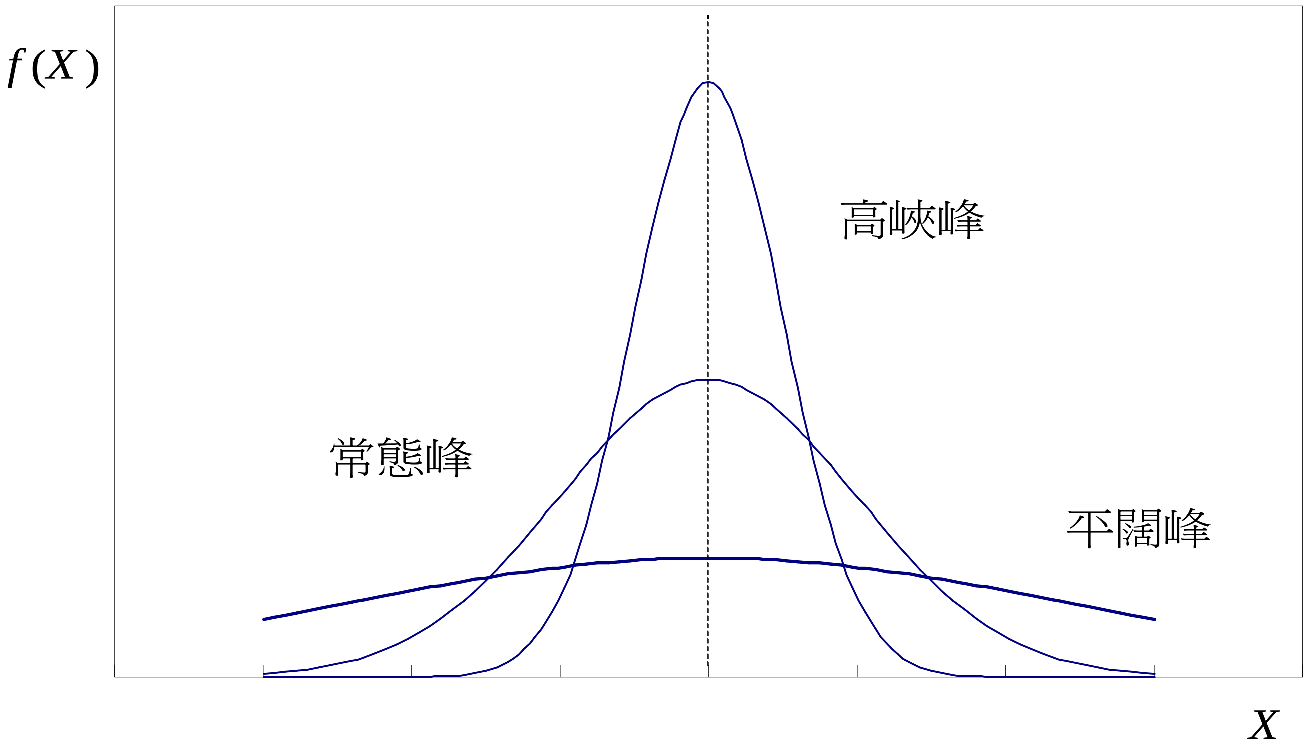
$$m_3 = \frac{1}{N}\sum(x_i - \bar{X})^3 = 2{,}618{,}226{,}082$$

$$\alpha_3 = \frac{m_3}{(\sqrt{m_2})^3} = \frac{2{,}618{,}226{,}082}{2{,}901{,}801{,}177} = 0.9023$$

右偏分配,表示少數的刷卡金額很大

4.5 未分組資料峰度的衡量

圖4.12 三種峰度的圖形



如何衡量峰度

☆ 峰度係數

母體:

$$\alpha_4 = \frac{M_4}{M_2^2} = \frac{M_4}{\sigma^4}$$

M_4為母體四級動差

樣本

$$\alpha_4 = \frac{m_4}{m_2^2}$$

m_4為母體四級動差

☆ 性質

1. 峰度一定為正數
2. 根據峰度係數的值可區分為
   a. $\alpha_4=3$為常態峰, $\alpha_4>3$為高狹峰, $\alpha_4<3$為平闊峰

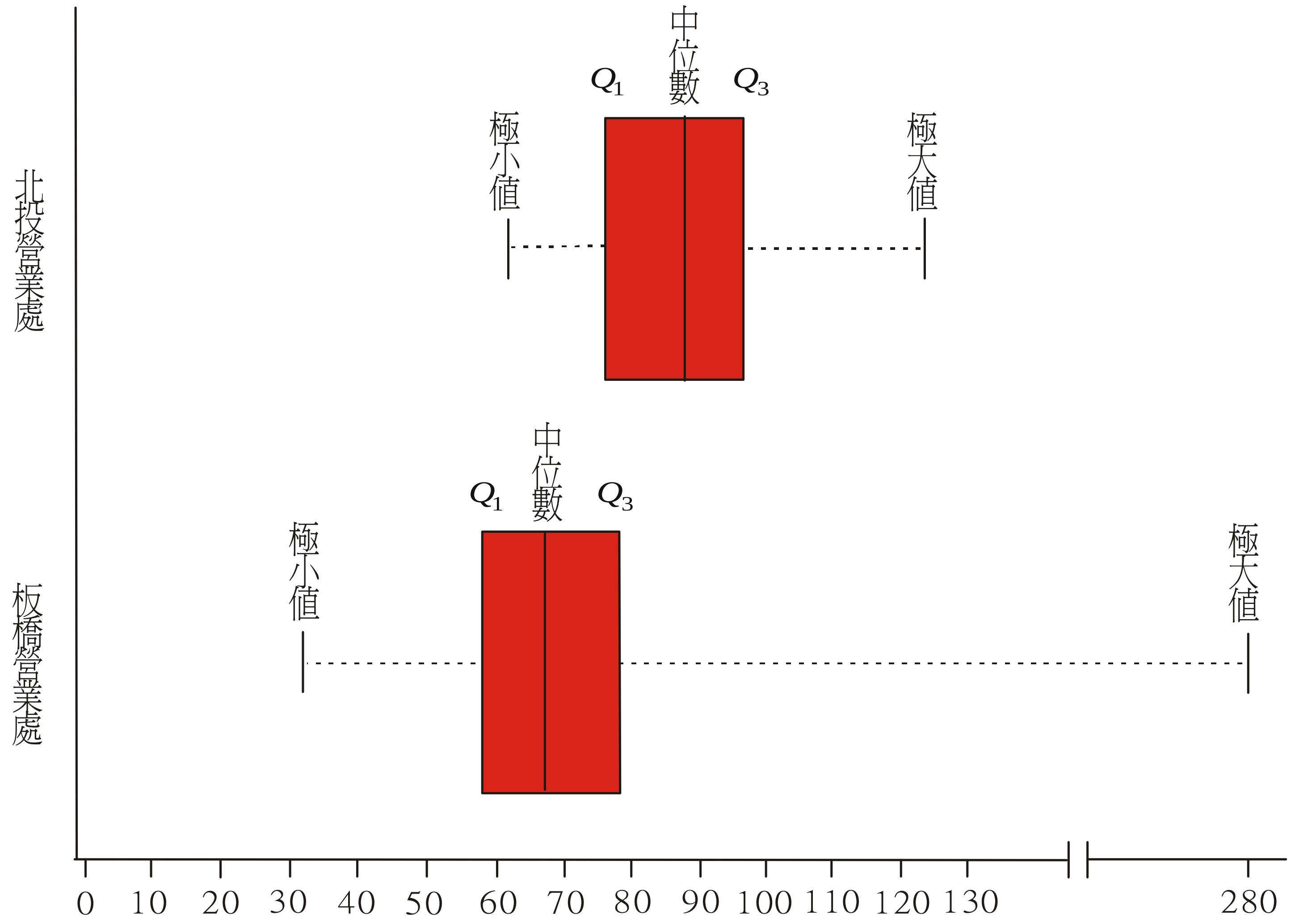
4.6 盒鬚圖分析法 (Box-and-whisker plot)

利用5個重要的測量數 Q_1, Me, Q_3, 最小值(min)和最大值(max)來完整的描述資料的性質

Ex 板橋及北投營業處的業績

板橋	32	52	64	64	70	76	82	280
北投	61	72	81	81	95	97	101	124

以板橋爲例 min=32 Q_1=58 Me=67 Q_3=79 max=280



林惠玲 陳正倉著 雙葉書廊發行 2004

說明:

1. 畫出盒子. 盒鬚圖係以第1及第3四分位數爲盒子的邊界,亦即50%的資料散佈在第一四分位數Q1=58與第三四分位數Q3=79之間
2. 在盒子中的中位數 (Me=67)劃一垂直線,該垂直線將此盒子中的資料分爲相等的兩部份(各25%),這是資料的中心位置
3. 從盒子左邊界線劃一虛線到最小值min=32,及從盒子的右邊界劃一虛線到最大值max=280,虛線就是所謂的鬍鬚(whisker),這代表資料的分散情形. 右邊的鬍鬚很長,表示有一超高的業績
4. 由盒子的寬窄可知居中50%資料的分散或集中情形.若盒子很寬表示50%資料分散度很大

4.7 分組資料中央趨勢的衡量

4.1 算術平均數

母體平均數: $\mu = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

樣本平均數 $\bar{X} = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

fi爲組次數 xi爲組中點 k爲組數

若有開放組距, 則取前一組兩倍之組距爲組距,在求組中點

求84檔股票型基金的平均投資報酬率

表4.8 股票型基金近三年報酬率的次數分配表

組別	組限（%）	組距（%）	組中點（%）	次數 f_i
1	$-45 < x \le -30$	15	−37.5	1
2	$-30 < x \le -15$	15	−22.5	5
3	$-15 < x \le 0$	15	−7.5	18
4	$0 < x \le 15$	15	7.5	21
5	$15 < x \le 30$	15	22.5	19
6	$30 < x \le 45$	15	37.5	12
7	$45 < x \le 60$	15	52.5	4
8	$60 < x \le 75$	15	67.5	4
				$\sum f_i = 84$

表4.9 股票型基金近三年報酬率的次數分配表

	A	B	C	D	E	F	G
1	組別	組限(%)	組距(%)	x_i 組中點(%)	f_i 次數	$f_i x_i$	累加次數
2	1	$-45 < x \le -30$	15	-37.5	1	-37.5	1
3	2	$-30 < x \le -15$	15	-22.5	5	-112.5	6
4	3	$-15 < x \le 0$	15	-7.5	18	-135	24
5	4	$0 < x \le 15$	15	7.5	21	157.5	45
6	5	$15 < x \le 30$	15	22.5	19	427.5	64
7	6	$30 < x \le 45$	15	37.5	12	450	76
8	7	$45 < x \le 60$	15	52.5	4	210	80
9	8	$60 < x \le 75$	15	67.5	4	270	84
10	合計				84	1230	

林惠玲 陳正倉著 雙葉書廊發行 2004

由公式可得

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1230}{84} = 14.6$$

4.1.2 中位數

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

式中 Lm_e: m_e所在組的組下界

Wm_e: m_e所在組的組距

fm_e: m_e所在組的組次數

F_L: m_e 前一組的累加次數

林惠玲 陳正倉著 雙葉書廊發行 2004

例4.21 股票型基金近三年報酬率的中位數

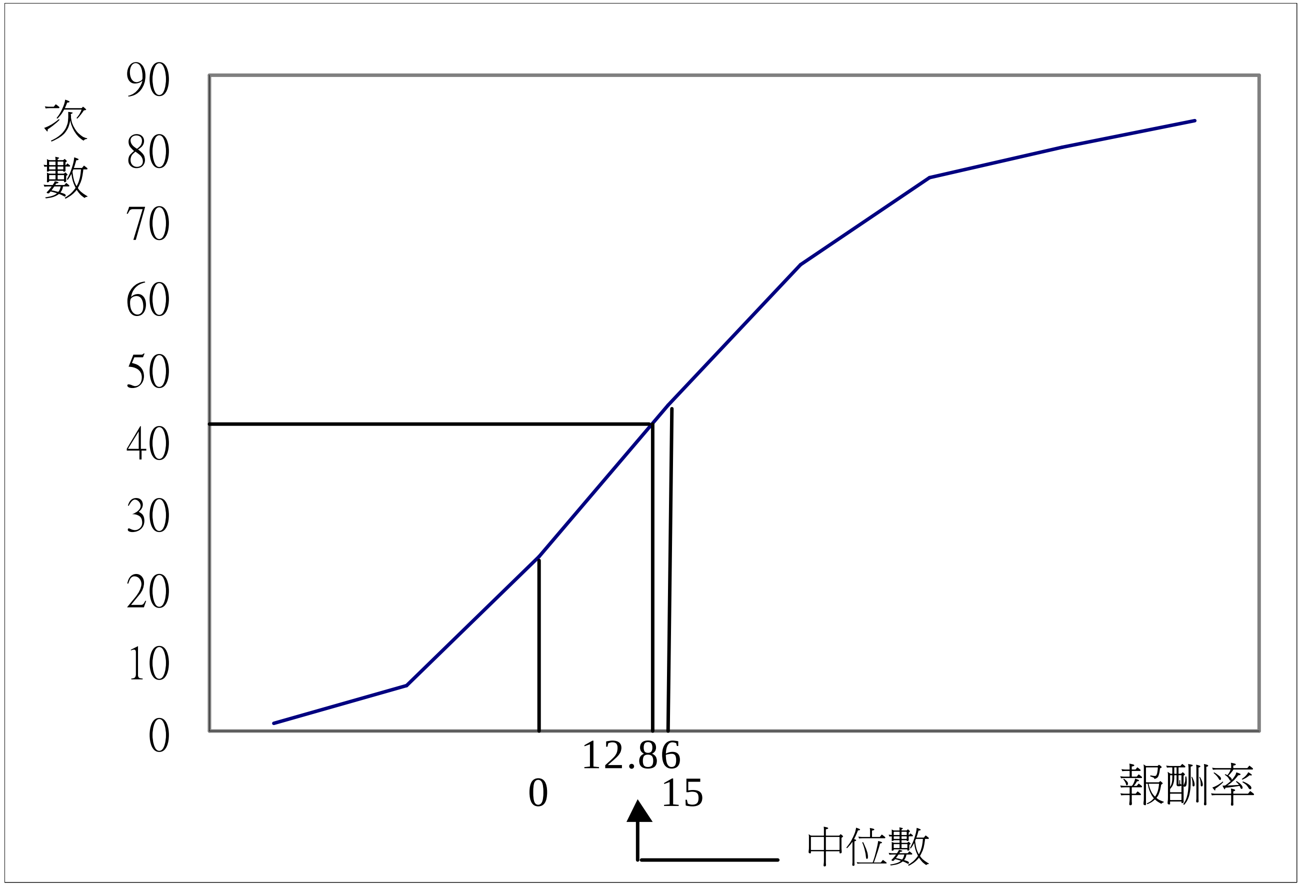
Solution:

中位數爲n/2=84/2=42個,位於第4組

$$m_e = L_{m_e} + W_{m_e}\left(\frac{\frac{n}{2} - F_L}{f_{m_e}}\right) = 0 + 15 \times \left(\frac{\frac{84}{2} - 24}{21}\right) = 0 + 12.86 = 12.86$$

林惠玲 陳正倉著 雙葉書廊發行 2004

圖4.14　股票基金近三年報酬率中位數的圖解



4.7.3 眾數

四種:

粗略法(crude method), 皮爾生法(Pearson's method)
金氏法(King's method), 克魯伯法(Czuber's method)

1.粗略法 (直接取組中點當眾數)

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

2. 皮爾生眾數

$$m_0 = \overline{X} - 3(\overline{X} - m_e)$$

例4.22 股票型基金近三年造酬率最多的是什麼

粗略法

(0+15)/2 =7.5 (7.5%)

皮爾生法

$$m_0 = \overline{X} - 3(\overline{X} - m_e)$$

=14.6-3(14.6-12.86)

=14.6-5.22

=9.38%

4.8 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。

例4.23 股票型基金近三年報酬率的變異

表4.10 一般股票型基金近三年報酬率的變異數與標準差

	A	B	C	D	E	F
1	組別	組限(%)	x_i 組中點(%)	f_i 次數	$x-\overline{X}$	$(x-\overline{X})^2 f_i$
2	1	$-45<x\leq-30$	-37.5	1	-52.1	2714.41
3	2	$-30<x\leq-15$	-22.5	5	-37.1	6882.05
4	3	$-15<x\leq0$	-7.5	18	-22.1	8791.38
5	4	$0<x\leq15$	7.5	21	-7.1	1058.61
6	5	$15<x\leq30$	22.5	19	7.9	1185.79
7	6	$30<x\leq45$	37.5	12	22.9	6292.92
8	7	$45<x\leq60$	52.5	4	37.9	5745.64
9	8	$60<x\leq75$	67.5	4	52.9	11193.64
10				$\Sigma=84$		43864.44

林惠玲 陳正倉著 雙葉書廊發行 2004

Solution:

由表4.10知

$$\sum (x_i - \overline{X})^2 \cdot f_i = 43864.44$$

故樣本變異數與標準差爲

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{8}(x_i - \overline{X})^2 \cdot f_i = \frac{43293.96}{(84-1)} = 528.487$$

$$S = \sqrt{S^2} = \sqrt{528.487} = 22.99$$

例4.24 求出級市場與次級市場相對利率的變化

表4.11 30天期商業本票的利率

組號	利率（%）	次數	
		初級市場	次級市場
1	$0 \le X < 2.0$	16	20
2	$2.0 \le X < 4.0$	16	13
3	$4.0 \le X < 6.0$	43	51
4	$6.0 \le X < 8.0$	36	34
5	$8.0 \le X < 10.0$	9	2

資料來源：EPS 台灣金融統計月報。

解:

初級市場的平均利率 μ_1=5.1%

次級市場的平均利率 μ_2=4.75%

變異數

$$\sigma_1^2 = \frac{1}{120}\left[16(1-5.1)^2 + 16(3-5.1)^2 + \ldots + 9(9-5.1)^2\right] = 5.057$$

$$\sigma_2^2 = \frac{1}{120}\left[20(1-4.75)^2 + 13(3-4.75)^2 + \ldots + 2(9-4.75)^2\right] = 4.438$$

變異係數爲

$$CV_1 = \frac{\sqrt{5.057}}{5.1} = 0.4409 \qquad CV_2 = \frac{\sqrt{4.438}}{4.75} = 0.4435$$

故次級市場的利率相對不穩定

林惠玲 陳正倉著 雙葉書廊發行 2004

4.9 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。

林惠玲　陳正倉著 雙葉書廊發行 2004

例4.25 試求學生英文成績的四分位數

表4.12 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

解:

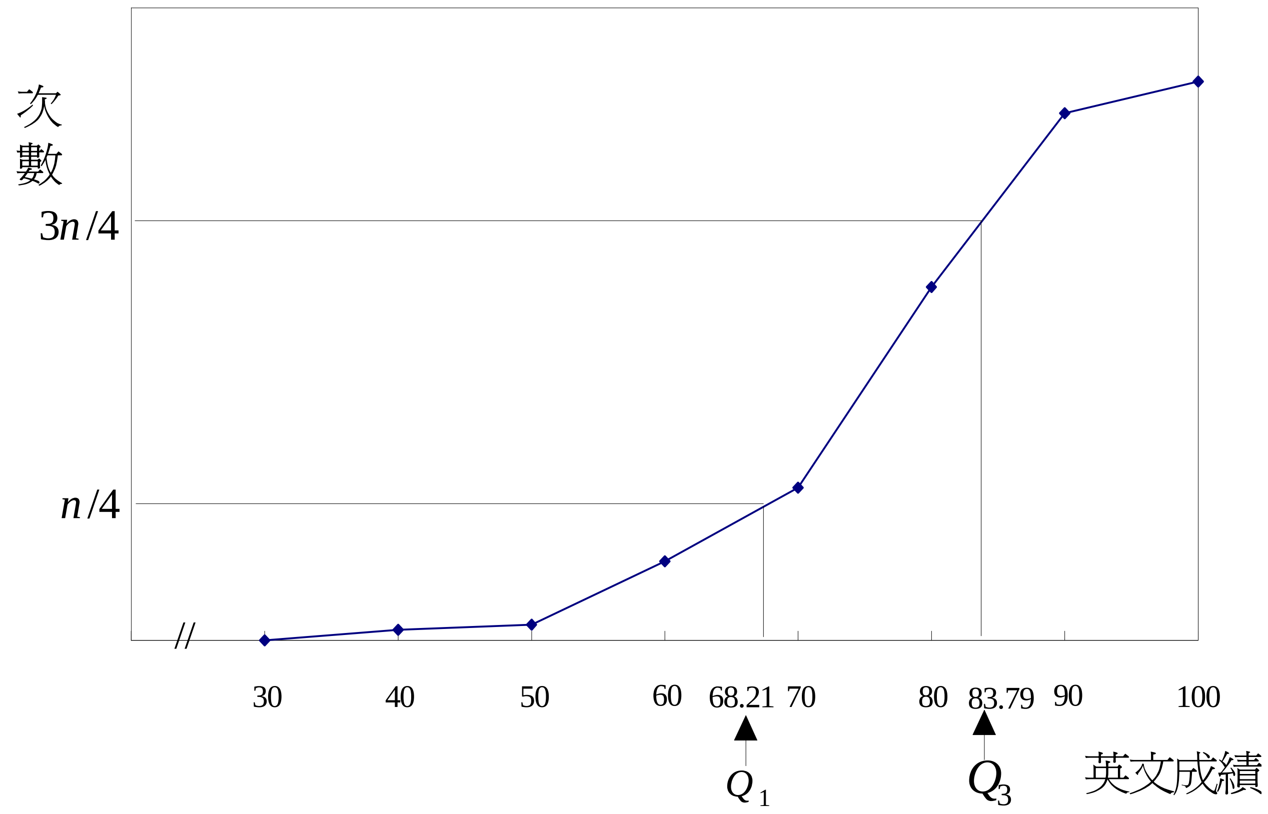
第一四分位數 (1/4)*106=26.5 落在第四組

$$Q_1 = L_{Q_1} + \frac{\frac{1}{4}n - F_{Q_1}}{f_{Q_1}} W_{Q_1} = 60 + \frac{26.5 - 15}{14} \times 10 = 68.21$$

第3四分位數 (3/4)*106=79.5 落在第6組

$$Q_3 = L_{Q_3} + \frac{\frac{3}{4}n - F_{Q_3}}{f_{Q_3}} W_{Q_3} = 80 + \frac{\frac{3}{4} \times 106 - 67}{33} \times 10 = 83.79$$

圖4.15 分組資料時，Q_1、Q_3的解法



4.9.2 十分位數

分組資料的十分位數

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

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。

4.9.3 百分位數

分組資料的百分位數

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

式中：Pi：第*i*個百分位數，L_{P_i}：Pi所在組的組下界

f_{P_i}：Pi所在組的組次數，

W_{P_i}：Pi所在組的組距

F_{P_i}：Pi前一組的累加次數。

例 4.26 and 4.27

Q: 已知大一英文成績如表4.12, a.求算第6個十分位數. b.某生想要成績在前5%,問應考幾分?

Solution:

a. 因 (6/10)*106=63.6,第6個十分位數落在第10組

$$D_6 = L_{D_6} + \frac{\frac{n\cdot 6}{10} - F_{D_6}}{f_{D_6}} W_{D_6} = 70 + \frac{\frac{106\times 6}{10} - 29}{38} \times 10 = 79.1$$

b. 第95百分位, 因 (95/100)*106=100.7, 落在第7組

$$P_i = L_{P_i} + \frac{\frac{n\cdot i}{10} - F_{P_i}}{f_{P_i}} W_{P_i} = 90 + \frac{\frac{106\times 95}{10} - 100}{6} \times 10 = 90 + 1.167 = 91.167$$

林惠玲 陳正倉著 雙葉書廊發行 2004

9.10 未分組資料的衡量-偏態與峰度

4.10.1 分組資料偏態的衡量

動差法的母體偏態係數

$$\alpha_3 = \frac{M_3}{\left(\sqrt{M_2}\right)^3} = \frac{\dfrac{\sum (x-\mu)^3 f_i}{N}}{\left(\sqrt{\dfrac{\sum (x-\mu)^2 f_i}{N}}\right)^3}$$

動差法的樣本偏態係數

$$\alpha_3 = \frac{m_3}{\left(\sqrt{m_2}\right)^3} = \frac{\dfrac{\sum (x-\overline{X})^3 f_i}{n}}{\left(\sqrt{\dfrac{\sum (x-\overline{X})^2 f_i}{n}}\right)^3}$$

例4.28 30天期商業本票利率分佈的偏度

如表4.11,計算出級市場與次級市場利率的偏度

Solution:

1. Pearson係數

以粗略法求眾數 $M_o^1=5,\ M_o^2=5$,由例4,24, $\mu_1=5.1$, $\mu_2=4.75$,
標準差分別為

$$\sigma_1=\sqrt{5.057}=2.248,\ \sigma_2=\sqrt{4.438}=2.107$$

因此可得Pearson偏態係數

$$SK_{P1}=\frac{5.1-5}{2.248}=0.0445,\ SK_{P2}=\frac{4.75-5}{2.107}=-0.1187$$

初級市場為右偏,次級市場為左偏, 次級市場呈現偏低

林惠玲 陳正倉著 雙葉書廊發行 2004

2. 動差法

初級市場與次級市場的三級動差分別爲

$$M_3^1 = \frac{1}{120}\left[(1-5.1)^3 \times 16 + \ldots + (9-5.1)^3 \times 9\right] = -3.918$$

$$M_3^2 = \frac{1}{120}\left[(1-4.75)^3 \times 20 + \ldots + (9-4.75)^3 \times 2\right] = -4.856$$

初級市場與次級市場的二級動差分別爲 $M_2^1 = 5.0567, M_2^2 = 4.4375$

所以偏態係數爲

$$\alpha_3^1 = \frac{-3.918}{\left(\sqrt{5.0567}\right)^3} = -0.3446 \qquad \alpha_3^2 = \frac{-4.856}{\left(\sqrt{4.4375}\right)^3} = -0.5195$$

初級與次級市場均有左偏,然而偏態並不顯著

動差法較精確

林惠玲 陳正倉著 雙葉書廊發行 2004

4.10.2 分組資料峰度的衡量

☆峰度係數

母體

$$\alpha_4 = \frac{M_4}{\sigma^4} = \frac{\dfrac{\sum (x-\mu)^4 f}{N}}{\left(\dfrac{\sum (x-\mu)^2 f}{N}\right)^2}$$

樣本

$$\alpha_4 = \frac{m_4}{m_2^2} = \frac{\dfrac{\sum (x-\overline{X})^4 f}{n}}{\left(\dfrac{\sum (x-\overline{X})^2 f}{n}\right)^2}$$

林惠玲 陳正倉著 雙葉書廊發行 2004

例4.29 30天期商業本票利率分佈的峰度

Solution:

初級市場與次級市場的四級動差分別為

$$M_4^1 = \frac{1}{120}\left[(1-5.1)^4 \times 6 + \ldots + (9-5.1)^4 \times 9\right] = 61.53$$

$$M_4^2 = \frac{1}{120}\left[(1-4.75)^4 \times 20 + \ldots + (9-4.75)^4 \times 2\right] = 46.676$$

初級市場與次級市場的二級動差分別為 $M_2^1 = 5.0567, M_2^2 = 4.437$

因此可得出級市場與次級市場的峰度細數分別為

$$\alpha_4^1 = \frac{61.53}{2.249^2} = 2.405 < 3 \qquad \alpha_4^2 = \frac{46.676}{2.107^2} = 2.368 < 3$$

平闊峰

林惠玲 陳正倉著 雙葉書廊發行 2004

4.11 Excel的使用

Steps: “工具”，“資料分析”，“敘述統計”

Note: 該表的結果與分組資料的結果有些差異

	A	B	C
1	股票型基金的報酬率		
2			
3	平均數	14.55511905	
4	標準誤	2.445867389	
5	中間值	13.87	中位數
6	眾數	#N/A	
7	標準差	22.4167449	
8	變異數	502.5104518	
9	峰度	-0.004619895	
10	偏態	0.364213066	
11	範圍	110.33	全距
12	個數	84	樣本數