



觀測量及其分析

回顧一些統計學上的概念

概述

- 資料集可利用圖形或數值方式來表達與分析
 - 如民調可用數值也可用圖形來表示
 - 又如股票的指數分析可用數值或圖形來進行

樣本VS.母數

■ 母數

- 是特定量的所有可能的觀測量
- 是一個無限大的量

■ 樣本

- 是由母數中選出的子集合
- 測量中的距離、角度或高程量測，實際上是由其母數中所取得的樣本

範圍與中值

- 範圍(range)又稱為分散度(dispersion)，是指樣本散佈的區域，即樣本最大值與最小值所涵蓋的區域
 - 如表2-1中，範圍=26.1-20.1=6.0
- 中值(median)是樣本按大小順序排列後，中間那個樣本的值
 - 若樣本總數為偶數，則取在中間的兩個樣本的平均值為中值
 - 如表2-2中，中值為23.4與23.5的平均值23.45

表2.1 五十個方向讀數

22.7	25.4	24.0	20.5	22.5
22.3	24.2	24.8	23.5	22.9
25.5	24.7	23.2	22.0	23.8
23.8	24.4	23.7	24.1	22.6
22.9	23.4	25.9	23.1	21.8
22.2	23.3	24.6	24.1	23.2
21.9	24.3	23.8	23.1	25.2
26.1	21.2	23.0	25.9	22.8
22.6	25.3	25.0	22.8	23.6
21.7	23.9	22.3	25.3	20.1

表2.2 排序後資料

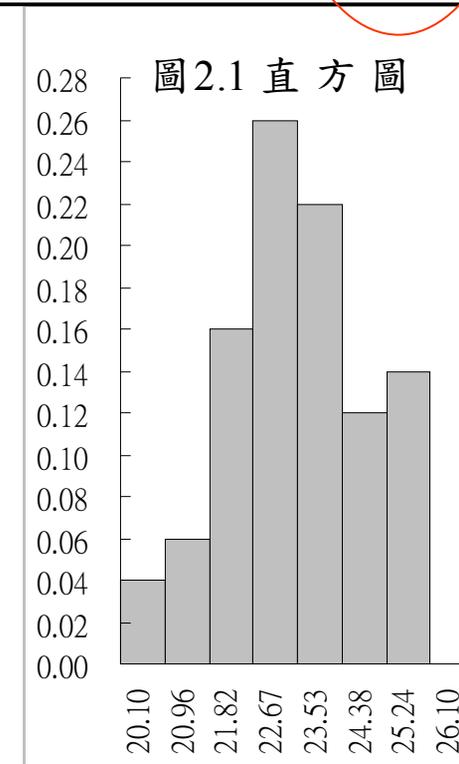
20.1	22.5	23.1	23.8	24.8
20.5	22.6	23.2	23.9	25.0
21.2	22.6	23.2	24.0	25.2
21.7	22.7	23.3	24.1	25.3
21.8	22.8	23.4	24.1	25.3
21.9	22.8	23.5	24.2	25.4
22.0	22.9	23.6	24.3	25.5
22.2	22.9	23.7	24.4	25.9
22.3	23.0	23.8	24.6	25.9
22.3	23.1	23.8	24.7	26.1

資料的圖形表達

- 直方圖(histogram)是表達資料頻率分佈(出現頻率)的條塊圖
- 直方圖的建立
 - 按照資料範圍，將資料分類(divided into classes)，即資料的子區域(subregions)
 - 假設分為7類，如表2.3的(1)所示
 - 統計每一類中的資料數量，即出現頻率
 - 如表2.3的(2)所示
 - 在直角坐標中，以橫軸代表分類，並按分類數量給定區間大小；縱軸代表出現頻率，並按最大頻率的值給定頻率分布的間隔
 - 按所統計的資料即可繪出直方圖
 - 如圖2.1所示
- 注意：表2.3中的(3)相對頻率的總和永遠為1

表2.3 頻率的計算

(1) 類區間	(2) 分類 頻率	(3) 分類相 對頻率
20.10~20.96	2	2/50 = 0.04
20.96~21.82	3	3/50 = 0.06
21.85~22.67	8	8/50 = 0.16
22.67~23.53	13	13/50 = 0.26
23.53~24.38	11	11/50 = 0.22
24.38~25.24	6	6/50 = 0.12
25.24~26.10	7	7/50 = 0.14
	$\Sigma = 50$	50/50 = 1



資料的圖形表達

■ 直方圖的功能

- 資料是否與中間值對稱
- 了解觀測值的範圍或分散度
- 了解觀測值的發生頻率
- 了解直方圖的陡峭度，即觀測值的精密度

■ 直方圖的形狀

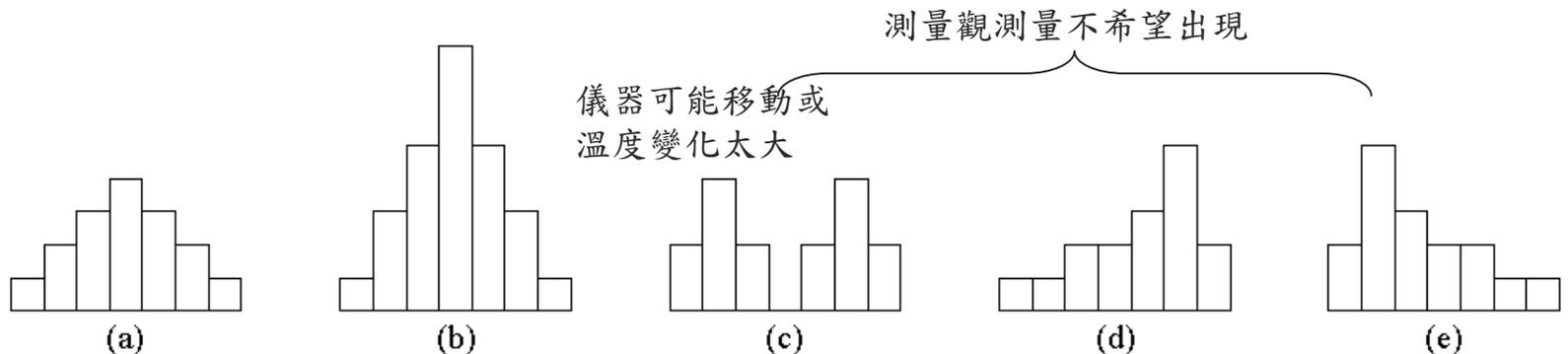


圖2.2 一般直方圖形狀

EDM測量期間穩定氣象
前緣出現或GPS的多路徑
效應

描述資料的數值方法

- 由資料集計算得到的數值描述子的值，常用來判讀資料的精密度或品質，此描述子有三類(均稱為統計)
 - 趨中度(central tendency)指標
 - 資料變化指標
 - 相對位置指標
- 所謂統計是一種由樣本資料計算而得的數值描述子

趨中度指標

- 是一種計算統計量
 - 是在資料集中趨向中央的一個數值表示
 - 算術平均值、中值、眾數與中數是四種這樣的指標
- 算術平均值為所有樣本(觀測量)加總的平均值
 - 如表2.2中之算術平均值為23.5
- 中值為樣本依序排列，位於中央的樣本值，若樣本為偶數，則區中央兩個樣本的平均值
 - 表2.2中之中值為23.45
- 眾數為在樣本中出現頻率最多的樣本值
 - 在測量中較少使用，因測量的樣本數不多
 - 表2.2中之眾數為23.8
- 中數為樣本中最大與最小樣本的平均值
 - 表2.2中之中數為23.1

$$\bar{y} = \sum_{i=1}^n y_i/n \quad (2.1)$$

其他定義

- 真值 μ ：為理論上的正確值(即母數平均值)，無法得知
- (真)誤差 ε ：任一觀測量與其真值的差，因真值無法得知，故(真)誤差也無法獲得
- 最或是值：在各觀測中該觀測值出現的機率最高
- 殘差 ν ：任一觀測量與其最或是值的差
- 自由度：超過求解未知數所需之觀測數量
 - 即多餘觀測之數量
 - 如量測兩點間距離三次，因觀測一次即可決定未知距離，所以其餘兩次即為多餘觀測
 - 多餘觀測可看出觀測值彼此之間的差異與不一致性
 - 重要的是有多於觀測才能進行平差工作，也才能獲得最或是值

其他定義

- 變異數(或變方)：為已知資料集中的精密度，或資料集中各個資料互相接近的程度，在統計學上有兩種變異數
 - 母數變異數：為誤差平方的平均值，更精確的說法為，誤差平方和除以其自由度
 - 樣本變異數：為殘差平方和除以其自由度
- 資料集中所有誤差之平均值不用來做為精密度之指標，因為偶然誤差有可能為正，也可能為負，且正負出現的機率相等，因此平均值必為0。
- 同理，殘差之平均值亦為0。

$$\begin{aligned}\sum_{i=1}^n \varepsilon_i &= \sum_{i=1}^n (y_i - \mu) = \sum_{i=1}^n y_i - \sum_{i=1}^n \mu = \sum_{i=1}^n y_i - n\mu = \sum_{i=1}^n y_i - n\left(\sum_{i=1}^n y_i / n\right) \\ &= \sum_{i=1}^n y_i - \sum_{i=1}^n y_i = 0\end{aligned}$$

其他定義

- 標準誤差：母數變異數的平方根
 - 無法求得
- 標準偏差(簡稱標準差)：樣本變異數的平方根
 - 測量中以此為衡量精度的指標
- 平均值之標準偏差：因為所有觀測值都含有誤差，因此，由觀測量的樣本集所計算之平均值也含有誤差
 - 當 $n \rightarrow \infty$ ，平均值之標準偏差 $\rightarrow 0$ 。此意謂當樣本集合趨近於整個母數時，計算之平均值亦趨近於真平均值 μ

$$\sigma = \pm \sqrt{\sum_{i=1}^n \varepsilon_i^2 / n}$$

$$S = \pm \sqrt{\sum_{i=1}^n v_i^2 / (n-1)}$$

$$S_{\bar{y}} = \pm S / \sqrt{n}$$

數例

例2.1 由表2.2的資料，利用式(2.7)與(2.10)，求樣本平均值、中值、眾數與標準偏差，並繪出直方圖(注意表2.2的資料為50個量測方向之秒位數)。

解 平均值：利用式(2.1)與表2.4的 Σy_i 值，可得

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^{50} y_i}{50} = \frac{1175}{50} = 23.5''$$

表2.4 例2.1之排序後資料

No.	y	v	v ²	No.	y	v	v ²	No.	y	v	v ²	No.	y	v	v ²	
1	20.1	3.4	11.56	14	22.7	0.8	0.64	27	23.6	-0.1	0.01	40	24.7	-1.2	1.44	
2	20.5	3.0	9.00	15	22.8	0.7	0.49	28	23.7	-0.2	0.04	41	24.8	-1.3	1.69	
3	21.2	2.3	5.29	16	22.8	0.7	0.49	29	23.8	-0.3	0.09	42	25.0	-1.5	2.25	
4	21.7	1.8	3.24	17	22.9	0.6	0.36	30	23.8	-0.3	0.09	43	25.2	-1.7	2.89	
5	21.8	1.7	2.89	18	22.9	0.6	0.36	31	23.8	-0.3	0.09	44	25.3	-1.8	3.24	
6	21.9	1.6	2.56	19	23.0	0.5	0.25	32	23.9	-0.4	0.16	45	25.3	-1.8	3.24	
7	22.0	1.5	2.25	20	23.1	0.4	0.16	33	24.0	-0.5	0.25	46	25.4	-1.9	3.61	
8	22.2	1.3	1.69	21	23.1	0.4	0.16	34	24.1	-0.6	0.36	47	25.5	-2.0	4.00	
9	22.3	1.2	1.44	22	23.2	0.3	0.09	35	24.1	-0.6	0.36	48	25.9	-2.4	5.76	
10	22.3	1.2	1.44	23	23.2	0.3	0.09	36	24.2	-0.7	0.49	49	25.9	-2.4	5.76	
11	22.5	1.0	1.00	24	23.3	0.2	0.04	37	24.3	-0.8	0.64	50	26.1	-2.6	6.76	
12	22.6	0.9	0.81	25	23.4	0.1	0.01	38	24.4	-0.9	0.81					
13	22.6	0.9	0.81	26	23.5	0.0	0.00	39	24.6	-1.1	1.21					
													Σ=	1175.0	0.0	92.36

中值：表列共有偶數個觀測量，資料的中點位於第25與第26個之間，即23.4與23.5，故取平均為23.45。

眾數：表列出現最多者為23.8，樣本內共出現三次。

範圍、類寬、直方圖：見第2.4節(圖解展示資料)，直方圖則如圖2.1所示。

標準偏差：表2.4中之 v 與 v^2 分別為各量測值之殘差[利用式(2.3)計算]及其平方，利用式(2.7)與表2.4中之 $\Sigma v^2 = 92.36$ ，可得：

若求表2.4中各 y 值之平方，再求其總和，得：

$$S = \pm \sqrt{\frac{\sum_{i=1}^{50} v_i^2}{n-1}} = \sqrt{\frac{92.36}{50-1}} = \pm 1.37''$$

$$\Sigma (y_i)^2 = 27,704.86$$

另利用式(2.10)，

$$S^2 = \frac{27,704.86 - 50(23.5)^2}{50-1} = \frac{92.36}{49} = 1.88$$

故得： $S = \pm \sqrt{1.88} = \pm 1.37''$

在表2.4中，與平均值相距一個標準偏差，即位於 $(23.5-1.37)$ 與 $(23.5+1.37)$ ，或位於22.13與24.87之間的量測共有34個，佔樣本中所有量測個數之 $34/50 \times 100\%$ ，或68%，符合前所提之理論。由表2.4也可看出：所有殘差之代數和為零。

圖2.1之直方圖繪出各類別值對應之相對頻率，由圖可見：數值群組趨近中央點，是相當精密的一組資料集。

例2.2 下列資料為50個方向量測之秒位數，求其平均值、中值、眾數，並利用式(2.10)求其標準偏差，並繪出直方圖。再與例2.1的資料比較。

34.2	33.6	35.2	30.1	38.4	34.0	30.2	34.1	37.7	36.4
37.9	33.0	33.5	35.9	35.9	32.4	39.3	32.2	32.8	36.3
35.3	32.6	34.1	35.6	33.7	39.2	35.1	33.4	34.9	32.6
36.7	34.8	36.4	33.7	36.1	34.8	36.7	30.0	35.3	34.4
33.7	34.1	37.8	38.7	33.6	32.6	34.7	34.7	36.8	31.8

解表2.5已將前列資料排序，並計算各量測值之平方；
 平均值：利用式(2.1)，求得

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{1737}{50} = 34.74''$$

中值：資料的中點位於第25與第26個之間，這兩個值均為34.7，故中值為34.7。

眾數：表列出現出現最多(三次)有三個不同數值，眾數即32.6、33.7、34.1。

表2.5 例2.2之排序後資料

No.	y	y ²												
1	30.0	900.00	12	33.4	1115.56	23	34.2	1169.64	34	35.6	1267.36	45	37.8	1428.84
2	30.1	906.01	13	33.5	1122.25	24	34.4	1183.36	35	35.9	1288.81	46	37.9	1436.41
3	30.2	912.04	14	33.6	1128.96	25	34.7	1204.09	36	35.9	1288.81	47	38.4	1474.56
4	31.8	1011.24	15	33.6	1128.96	26	34.7	1204.09	37	36.1	1303.21	48	38.7	1497.69
5	32.2	1036.84	16	33.7	1135.69	27	34.8	1211.04	38	36.3	1317.69	49	39.2	1536.64
6	32.4	1049.76	17	33.7	1135.69	28	34.8	1211.04	39	36.4	1324.96	50	39.3	1544.49
7	32.6	1062.76	18	33.7	1135.69	29	34.9	1218.01	40	36.4	1324.96			
8	32.6	1062.76	19	34.0	1156.00	30	35.1	1232.01	41	36.7	1346.89			
9	32.6	1062.76	20	34.1	1162.81	31	35.2	1239.04	42	36.7	1346.89			
10	32.8	1075.84	21	34.1	1162.81	32	35.3	1246.09	43	36.8	1354.24			
11	33.0	1089.00	22	34.1	1162.81	33	35.3	1246.09	44	37.7	1421.29			
													Σ=	1737.0 60584.48

表2.6 例2.2之頻率表

分類	分類 頻率	分類相 對頻率
29.15~30.01	1	0.02
30.01~30.87	2	0.04
30.87~31.73	0	0.00
31.73~32.59	3	0.06
32.59~33.45	6	0.12
33.45~34.31	11	0.22
34.31~35.17	7	0.14
35.17~36.03	6	0.12
36.03~36.89	7	0.14
36.89~37.75	1	0.02
37.75~38.61	3	0.06
38.61~39.47	3	0.06
	$\Sigma = 50$	$\Sigma = 1.00$

類寬：為與上例比較，類寬採用與表2.2相同之0.86，又因預期直方圖中央會趨近資料之平均值，故最中間之區間為平均值加減類寬之半(0.43)，故中央區間為自34.31(=34.74-0.43)至35.17(=34.74+0.43)，其他之分類區間則由中央區間之邊界向兩側減或加類寬，直到所有資料都在區間之邊界內，譬如中央區間之左側區間為33.45(=34.31-0.86)至34.31，中央區間之右側區間則為35.17至36.03(=35.17+0.86)，其餘分類區間仿此類推；分類頻率表如表2.6所列，而直方圖2.3也因此繪得。

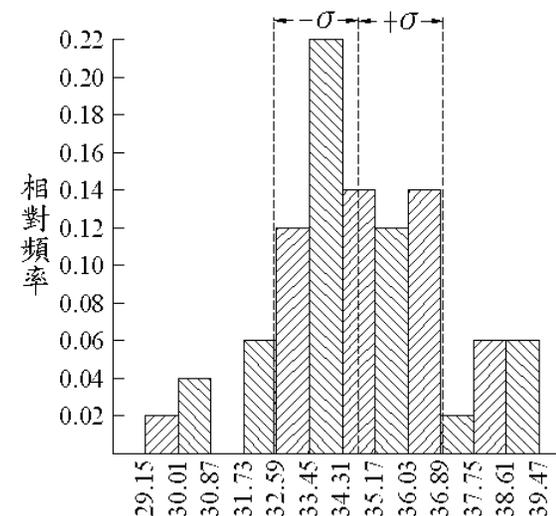


圖2.3 例2.2之直方圖

變方：根據式(2.10)，利用表2.5中所有量測平方的和，求得樣本之變方為：

$$S^2 = \frac{\sum_{i=1}^{50} y_i^2 - n\bar{y}^2}{n-1} = \frac{60584.48 - 50 \times 34.74^2}{50-1} = 4.92$$

樣本標準偏差： $S = \pm\sqrt{4.92} = \pm 2.22$

落於平均值 $\pm S$ 之邊界內(即 37.74 ± 2.22)的量測有30個，也就是全部量測之60%。接近68.3%之理論值，平均值及其加減一個標準差之邊界如圖2.3中之虛線所示。

比較：例2.2之標準差(± 2.22)大於例2.1之標準差(± 1.37)，例2.2之資料範圍(9.3)亦較例2.1者(6.0)為大，因此，例2.2之精度較例2.1之精度差；比較兩例之直方圖，亦可見其差異，圖2.1中央之寬度較圖2.3者窄又高。

□ 樣本變方的推導

如前所提，求母數之變方時，分母為 n ，求樣本之變方時，分母則為 $n-1$ ；所以有此不同，簡言之， n 個觀測量，其中一個需用來計算平均值，剩下 $n-1$ 個觀測量是用來計算變方。推導式(2.5)如下，可瞭解這些。

假設母數之平均值為 μ ，標準誤差為 σ ，從中選出 n 個樣本， y_i 為其中一個觀測量，則：

$$y_i - \mu = (y_i - \bar{y}) + (\bar{y} - \mu) = (y_i - \bar{y}) + \varepsilon$$

其中， $\varepsilon = \bar{y} - \mu$ 為樣本平均值之誤差，平方上式兩邊：

$$(y_i - \mu)^2 = (y_i - \bar{y})^2 + \varepsilon^2 + 2\varepsilon(y_i - \bar{y})$$

從 1 至 n ，加總樣本內所有觀測量：

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n\varepsilon^2 + 2\varepsilon \sum_{i=1}^n (y_i - \bar{y})$$

根據樣本平均值 \bar{y} 之定義：

$$\sum_{i=1}^n (y_i - \bar{y}) = \sum_{i=1}^n y_i - n\bar{y} = \sum_{i=1}^n y_i - \sum_{i=1}^n y_i = 0, \text{ 故：}$$

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n\varepsilon^2 + 2\varepsilon \sum_{i=1}^n (y_i - \bar{y}) = \sum_{i=1}^n (y_i - \bar{y})^2 + n\varepsilon^2$$

重複計算眾多樣本，上式左邊之平均值趨向 $n\sigma^2$ ；類似的，因為 ε 為樣本平均與母數平均之間的差異，故

$n\varepsilon^2 = n(\mu - \bar{y})^2$ 之平均值趨向 n 乘上 \bar{y} 之變方。因此，

$n\varepsilon^2 \rightarrow n(\sigma^2/n)$ ，式中， σ^2/n 為 $n \rightarrow \infty$ 時， \bar{y} 之變方。由上討論，可得：

$$n\sigma^2 \rightarrow \sum_{i=1}^n (y_i - \bar{y})^2 + \sigma^2 \text{ 或 } \sum_{i=1}^n (y_i - \bar{y})^2 \rightarrow (n-1)\sigma^2$$

因此，推得：

$$S^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1) \rightarrow \sigma^2$$

換言之，對個數很大之隨機樣本而言，上式之平均值趨向 σ^2 ；也就是： S^2 是母數變方之無偏差估值。