

第二章

統計資料之描述、陳示及探討

1

這種包括統計程序中的資料蒐集、描述及彙整資料的結果，乃為敘述統計（Descriptive Statistics）的基本要件。

2.1 The Frequency Distribution

我們知道統計學主要的目的：

- ◎母體或樣本的特性經由調查、實驗或研究中所獲得的資訊，以文字或數據表示出來便成了資料（Data）。
- ◎未經分類與整理的資料稱為原始資料（Raw Data）。
- ◎將原始資料依類別而分成若干組，將連續性資料劃分為若干的區段，稱之為組距（Class Intervals）。

2

- ◎將每一個觀測值計入所屬的組距內，再計算各組的次數，稱為組次數（Class Frequency）。而將資料依數量大小或類別而分成若干組並計算各組資料的個數（發生次數），以顯示資料分佈的過程，稱為樣本次數分配（Sample Frequency Distribution）。

3

表 2-1 試驗裝置校正的所需時間（秒）

12.8	15.6	13.5	15.7	15.3	15.2	20.1	14.2	12.9	14.0
16.9	14.3	15.5	14.6	13.0	14.7	19.0	13.0	11.3	14.2
14.5	14.8	14.2	13.0	13.1	12.5	16.1	19.1	16.7	13.2
15.0	12.7	13.6	13.3	13.2	14.7	12.9	13.1	17.3	15.4
17.9	13.0	14.3	14.2	15.7	15.6	13.0	13.9	14.2	16.0
12.9	13.1	13.3	12.3	13.1	13.6	13.2	18.5	13.2	13.7
12.6	14.4	14.5	13.9	17.0	13.7	12.7	16.8	13.3	14.7
14.2	13.0	14.6	14.0	12.9	14.7	12.8	12.0	14.2	12.8
13.7	15.2	14.8	13.0	11.7	12.2	13.3	13.8	14.2	14.3
14.7	12.6	18.9	14.3	14.4	15.5	16.8	17.0	13.2	12.9

4

表 2-2 檢查時間之樣本次數分配

時間 (秒)	記 錄	檢 查 次 數
11.0 - 小於 12.0	//	2
12.0 - 小於 13.0	//// //	16
13.0 - 小於 14.0	//// // // // //	29
14.0 - 小於 15.0	//// // // // // //	27
15.0 - 小於 16.0	//// // // /	11
16.0 - 小於 17.0	//// /	6
17.0 - 小於 18.0	//// /	4
18.0 - 小於 19.0	//	2
19.0 - 小於 20.0	//	2
20.0 - 小於 21.0	/	1
		總次數 100

5

2.1.1 直方圖與次數曲線(Histogram and Frequency Curve)

次數分配 (Frequency Distribution) 若以圖形表示, 則可使樣本資料更易於分析與結論。圖2-1所示為直方圖(Histogram)。其中縱軸為檢查次數之矩形圖, 橫軸為檢查時間的調查資料。

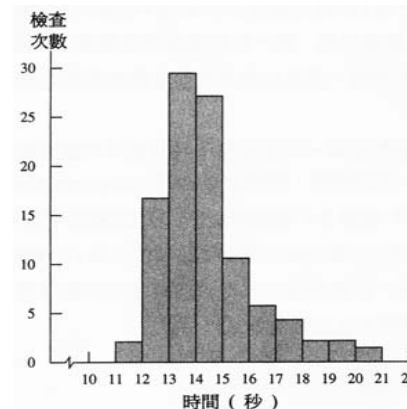


圖 2-1 檢查時間次數分配之直方圖

6

另外一種資料的圖形顯示的方式係為**次數多邊形圖 (Frequency Polygon)**, 如圖2-2係將直方圖資料以次數多邊形圖表示。

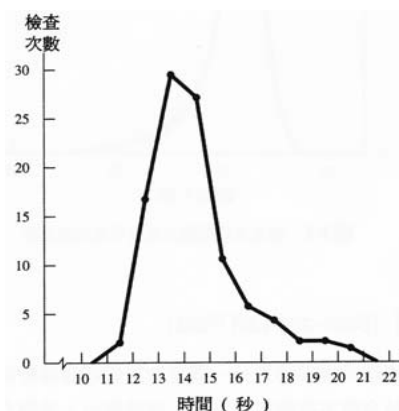


圖 2-2 檢查時間次數分配之多邊形圖

7

由於多邊形圖為一折線圖而不易判定其次數分佈函數的型式, 若將其繪修成一平滑曲線, 稱為**次數曲線 (Frequency Curve)** 如圖2-3

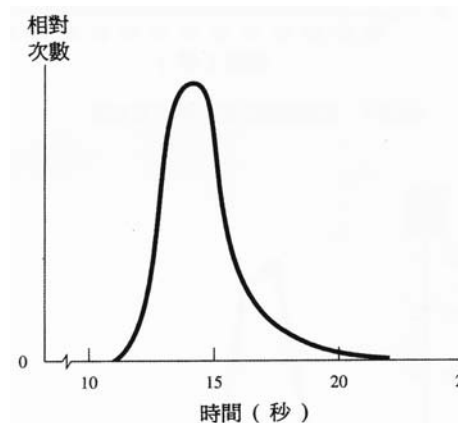


圖 2-3 檢查次數母體之建議平滑次數曲線

8

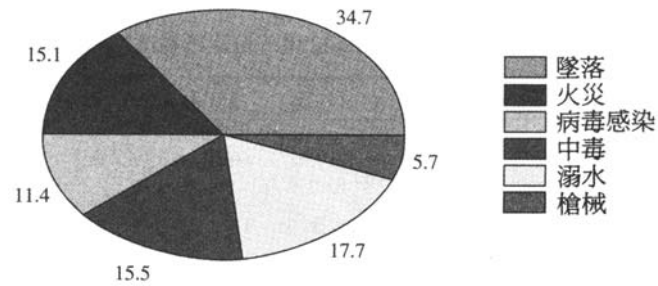


圖 2-6 意外死亡原因之圓形圖

2.1.2 相對次數與累積次數分配(Relative and cumulative Frequency Distribution)

由原始的次數分配中，每一組的次數除以其樣本數（或所有觀測值的個數）或稱為**相對次數分配（Relative Frequency Distribution）**。

表 2-4 樣本檢驗時間之相對次數分配

時間 (秒)	出現次數	相對次數	時間 (秒)	出現次數	相對次數
11.0-小於 12.0	2	$\frac{2}{100} = 0.02$	16.0-小於 17.0	6	$\frac{6}{100} = 0.06$
12.0-小於 13.0	16	$\frac{16}{100} = 0.16$	17.0-小於 18.0	4	$\frac{4}{100} = 0.04$
13.0-小於 14.0	29	$\frac{29}{100} = 0.29$	18.0-小於 19.0	2	$\frac{2}{100} = 0.02$
14.0-小於 15.0	27	$\frac{27}{100} = 0.27$	19.0-小於 20.0	2	$\frac{1}{100} = 0.01$
15.0-小於 16.0	11	$\frac{11}{100} = 0.11$	20.0-小於 21.0	1	Total 1.00

在許多情形下，我們所感興趣的可能不是落在某一組內的觀測個數，而是落在某一特定值之上或之下觀測值之個數，此時可使用**累積次數分配（Cumulative Frequency Distribution）**來加以分析。

表 2-5 檢驗時間的累積次數分配

查驗時間 (秒)	次數	累積次數	累積的相對次數
11.0-小於 12.0	2	2	0.02
12.0-小於 13.0	16	2 + 16 = 18	0.18
13.0-小於 14.0	29	18 + 29 = 47	0.47
14.0-小於 15.0	27	47 + 27 = 74	0.74
15.0-小於 16.0	11	74 + 11 = 85	0.85
16.0-小於 17.0	6	85 + 6 = 91	0.91
17.0-小於 18.0	4	91 + 4 = 95	0.95
18.0-小於 19.0	2	95 + 2 = 97	0.97
19.0-小於 20.0	2	97 + 2 = 99	0.99
20.0-小於 21.0	1	99 + 1 = 100	1.00

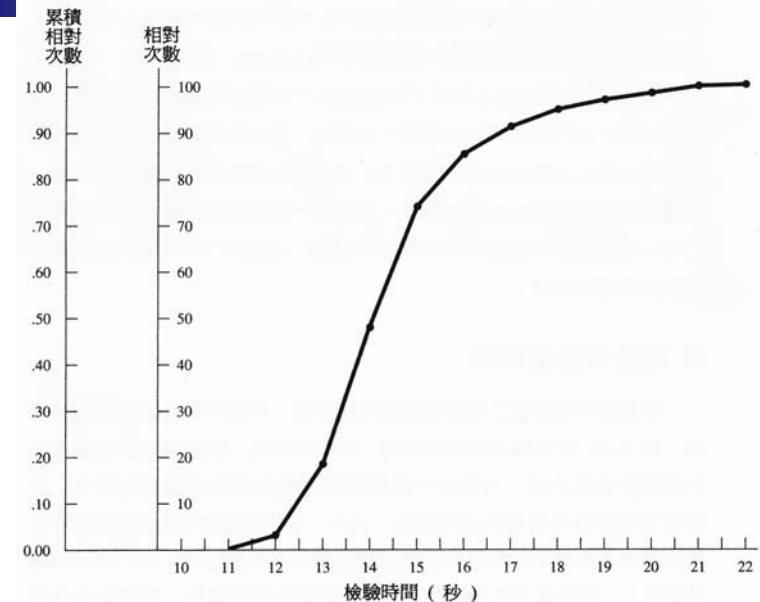
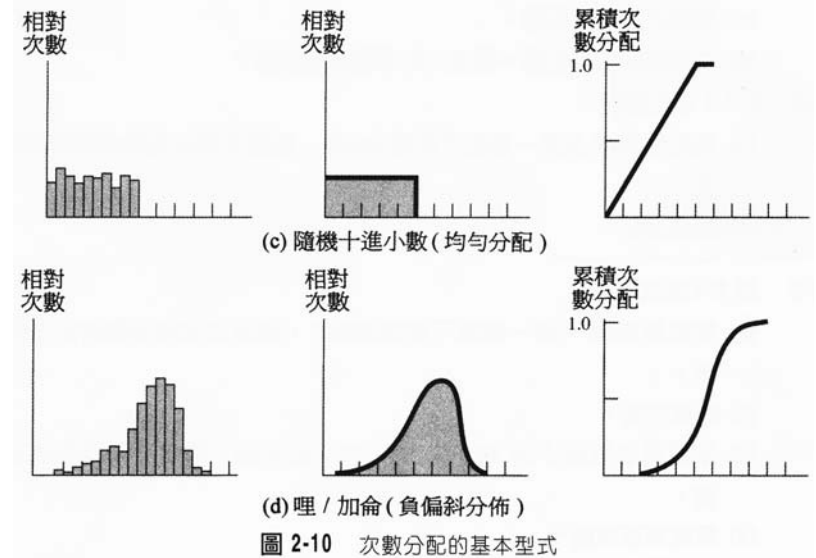
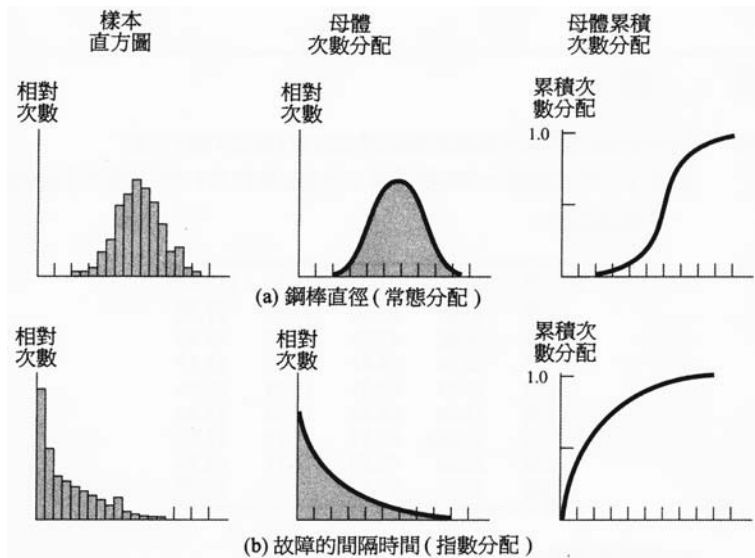


圖 2-7 檢驗時間的累積次數分配

2.1.3 次數分配的形狀(Common Forms of the Frequency Distribution)

計量樣本或母體之相對或累積次數分佈，可依不同的形狀而加以分類。圖2-10所示為次數分配圖的一些基本型式，由圖最左側之樣本直方圖的形狀及分佈，可用以作為推知其母體次數分佈型式之參考。

1. 圖2-10(a)表為鋼棒直徑之觀測資料，直方圖近似於鐘形分配曲線，而鐘形分配曲線屬於常態分配 (Normal Distribution)。
2. 常態分配為一對稱而無偏斜的曲線，然而許多資料呈現的是非對稱之偏斜曲線。非對稱而偏斜的次數分配圖的特徵是在圖形的右邊特別細長或左邊特別細長，前者稱為正偏分佈 (Positive Skewed Distribution)，而後者稱負偏分佈 (Negative Skewed Distribution)。



母體次數分配具有特定的數學函數型式（如常態分配與指數分配等）與參數值（如平均數與變異數等），故由樣本直方圖找出配適的母體分配型態後，則可據此分析樣本的特性。

17

2.2 Summary Statistical Measures : Location(位置的測度)

- 舉例來說，某位工業工程師欲在兩種不同生產方式中選擇其一生產較快速者，利用次數分配表或直方圖並無法清楚及容易地用以評量何種生產方式較快。
- 但若改以比較兩者生產完成時間平均值的差異，則能清楚地指出何種生產方法較快。但若改以比較兩者生產完成時間平均值的差異，則能清楚地指出何種生產方法較快。

18

1. 為何需要統計量數？

期望由次數分配中找出幾個容易表現一個次數分配的統計。

2. 統計量數有哪些？

(1) 資料的位置 (Location)。

量度資料的位置(Location)有兩種常用的方式：

- ◎ 第一種是以**集中趨勢**或稱**中央趨勢(Central Tendency)**來代表資料的中心位置（即中心點）的數值，此乃基於大部份的資料組，均會對其中心呈現出一明顯的趨勢的特性。

19

◎ 另一種則是以資料發生之次數分配的位置來量測各資料之相對不同位置 (Positions)。

(2) 評量各資料與其中心位置的差異程度來量測，即資料的**變異性 (Variability)** 或 **離散程度 (Dispersion)**。

上述為兩種最主要之統計量

20

2.2.1 統計量與參數(Statistics and Parameters)

用以衡量母體的量數或特徵值，稱為母體參數，

(Population Parameter)： μ ， σ

分析樣本資料的特徵（量數），此稱為**樣本統計量**

(Sample Statistic)：

21

2.2.2 算數平均數 (Arithmetic Mean)

X_i 若代表一筆原始資料中的第*i*個觀察值，樣本大小 (Sample Size) 為 n ，即共包含 n 個數值： X_1, X_2, \dots, X_n ，則此**樣本平均數**為

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

若以 f_k 代表第 k 組的組次數， X_k 為第 k 組的組中點 (Midpoint)，則分組資料之平均數的計算公式如下：其中 n 為各組次數之和，即 $n = \sum f_k$

$$\bar{X} = \frac{\sum f_k X_k}{n}$$

22

表 2-6 分組資料之平均數的計算表列

組界 (時間)	次數 (f_k)	組中點 (X_k)	$f_k X_k$
11.0-小於 12.0	2	11.5	23.0
12.0-小於 13.0	16	12.5	200.0
13.0-小於 14.0	29	13.5	391.5
14.0-小於 15.0	27	14.5	391.5
15.0-小於 16.0	11	15.5	170.5
16.0-小於 17.0	6	16.5	99.0
17.0-小於 18.0	4	17.5	70.0
18.0-小於 19.0	2	18.5	37.0
19.0-小於 20.0	2	19.5	39.0
20.0-小於 21.0	1	20.5	20.5

總和 1,442.0

$$\bar{X} = \frac{\sum f_k X_k}{n} = \frac{1,442.0}{100} = 14.42(\text{秒})$$

23

2.2.3 中位數 (Median)

集中趨勢的第二種量度方式為中位數 (Median)。當一組資料次數分配曲線呈偏斜時 (此時資料可能存有非常大或非常小的極端值)，則平均數的代表性將受質疑，此時採用中位數乃是最佳的量度。由於母體通常很大，故

◎一般所求得的中位數為**樣本中位數 (Sample Median)**，以符號 m 表示。

◎中位數係將資料按大小順序 (一般是由小依序排至大) 後，位於最中間的數的數值稱之。

24

◎當資料個數為奇數時，最中間數為其中位數；當資料個數為偶數時，則取最中間兩數的平均數為其中位數。
例如：

584 613 622 693 755

則樣本中位數 $m = 622$

由於中位數係資料**中央位置 (Position)**的平均數，並不受極大或極小之極端值的影響。

平均值與中位數應用時機？

1. 當資料個數 n 很大時，用平均數。
2. 資料有極大或極小之極端值時用中位數。

政府引用家庭所得資料通常是以使用中位數為準，而不使用平均數。

25

2.2.4 眾數 (The Mode)

第三種衡量集中趨勢的方法為眾數(Mode)。
眾數為資料中出現次數最多的數值然而眾數亦可能不只一個或不存在。

當資料為分組資料時，眾數乃為次數分配最多的那一組之組中點值。

26

例如：以下分組資料中眾數位於第2組組中央，即
 $(100.0+105.0) / 2 = 102.5$

組界	次數
95.0—小於 100.0	7
100.0—小於 105.0	23
105.0—小於 110.0	22
110.0—小於 115.0	17
115.0—小於 120.0	4
	$n = 73$

眾數使用時機：

當資料是以相對次數分配曲線表示時，以眾數來量度資料的集中趨勢，將較其他的量數更為有用。

27

2.2.5 次數分配的型式與總結量數(Frequency Distribution forms and Summary Measures)

1. 圖(a)為**對稱 (Symmetrical)**的分配曲線，代表平均數，中位數及眾數三者之值及位置皆為相同。
2. 呈現**偏態 (Skewed)**分配，則此三者衡量所在的位置將不同。
3. 在圖(b)為左偏分配；在圖(c)為右偏分配。

28

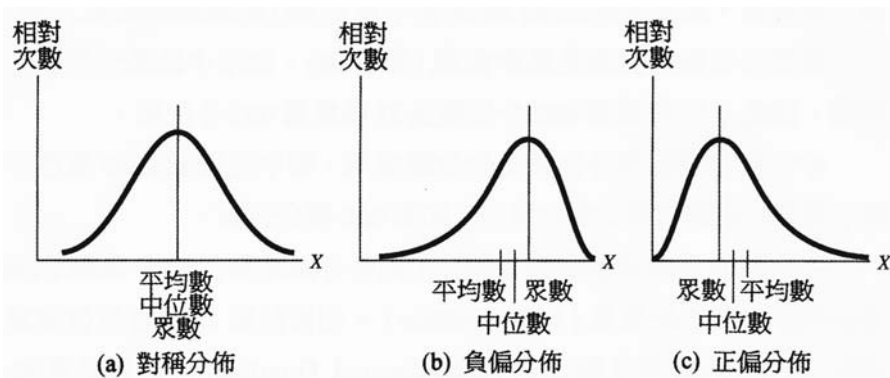


圖 2-12 三個集中趨勢量數在對稱與偏斜次數分配下其位置的比較

4. 當母體的次數分配出現2個峰狀，稱為**雙眾數分配**（**Bimodal Distribution**）。此時可利用眾數的觀念對母體的特性加以描述。

如圖2-13(a)代表男女學生身高出現之雙眾數分配的情形。此時最好將男性與女性學生的身高資料，加以區分其性別並單獨分析。通常雙眾數分配的發生，是因為原始資料具有異質性。

如圖2-13(b)顯示了某校兼職學生與全職學生其平均成績的雙峰分配，同樣地，我們應將兼職學生與全職學生的資料作個別的衡量。

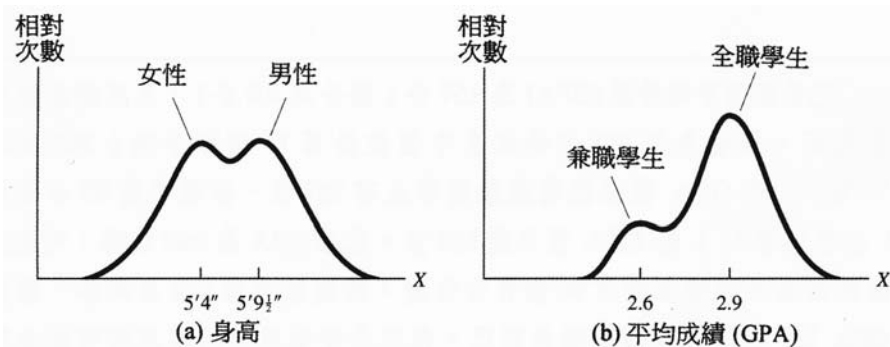


圖 2-13 雙眾數之次數分配

2.2.6 百分位數、分位數與四分位數

1. **百分位數 (Percentile)** 是另一種測度位的量數，係表示小於某特定百分點以下的量測值。百分位數的意義乃將資料由小至大排列後，分割成相同的100等分，而每一等分點皆百分位數。
2. **分位數 (Fractile)**，表示小於某分位點之量測值。
3. 若將資料分隔成相等的4等分，則各分隔點稱為**四分位數 (Quartile)**。**第1個四分位數 (First Quartile)**，相當於第25個百分位數或第0.25個分位數；**第2個四分位數 (Second Quartile)**，相當於第50個百分位數或第0.5個分位數；**第3個四分位數 (Third Quartile)**，則相於第75個百分位數或第0.75個分位數。

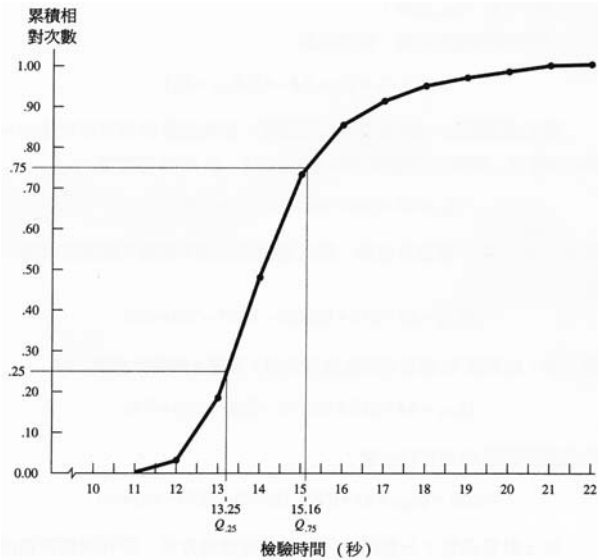


圖 2-14 由累積次數分配圖找出百分位數的示意圖

33

2.3 差異性的量度 (變異量數)

為什麼需要差異性量度：

由於我們所蒐集到的資料或多或少都有差異時存在，故除了資料之位置外，資料的變異量數 (Measures of Variability) 為另一種極為重要的敘述性的彙總。

主要是用於衡量一組資料中，各個觀測值之間的差異或離散的程度，並用以反映平均數代表性的強弱。

34

2.3.1 變異度的重要性

變異度 (Variability) 與離散度 (Dispersion) 為同義字，係用於描述各個觀測值間的差異或離散的程度。當各觀測值之變異度愈大，表示其離散的程度愈大。

35

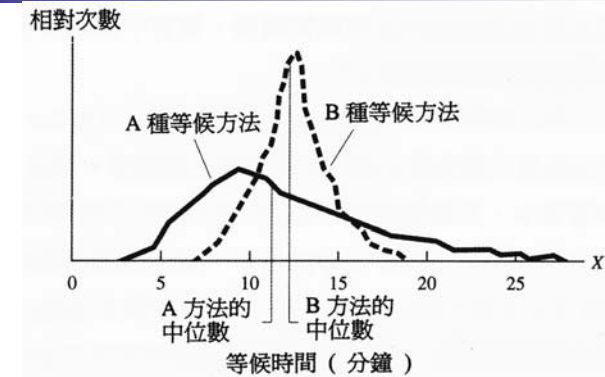


圖 2-15 兩種不同等候上機方法其等候時間之次數分配

討論：這個例子顯示出當以集中趨勢的量數無法充分地提供決策的參考時，變異度為另一種比較資料差異性的有用的統計分析方法。

36

2.3.2 全距 (Range)

- ◎一組資料中，數值最大者與最小者之差稱為**全距**。一般以 R 表示。
- ◎全距最主要之缺點為易受**極端值**或稱**界外值 (Outlier)**的影響，而且無法得知除最大值與最小值外之資料的差異情形。

2.3.3 四分位距與箱形圖

四分位距 (Interquartile Range) 可改善以全距測度異度或離散度之缺點，改用資料的第3個四分位數 ($Q_{0.75}$) 與第1個四分位數 ($Q_{0.25}$) 之差，來代表資料中間一半的觀測值之全距。則四分位距為 $Q_{0.75} - Q_{0.25}$

結合全距與四分位距可繪製成**箱形圖 (Box Plot)**

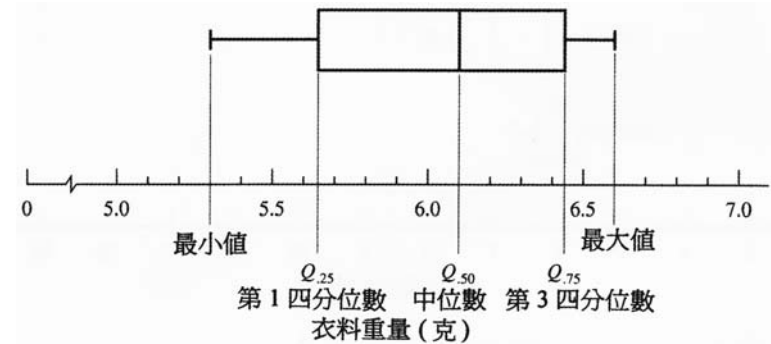


圖 2-16 衣料重量樣本之箱形圖

2.3.4 變異數與標準差

- ◎資料變異性最重要的測度乃基於各觀察值與其集中量數之差異，而此差異稱之為**離差 (Deviation)**。
- ◎資料中各個觀測值與其平均數之離差有正亦有負，然而各個離差的總和必為0。

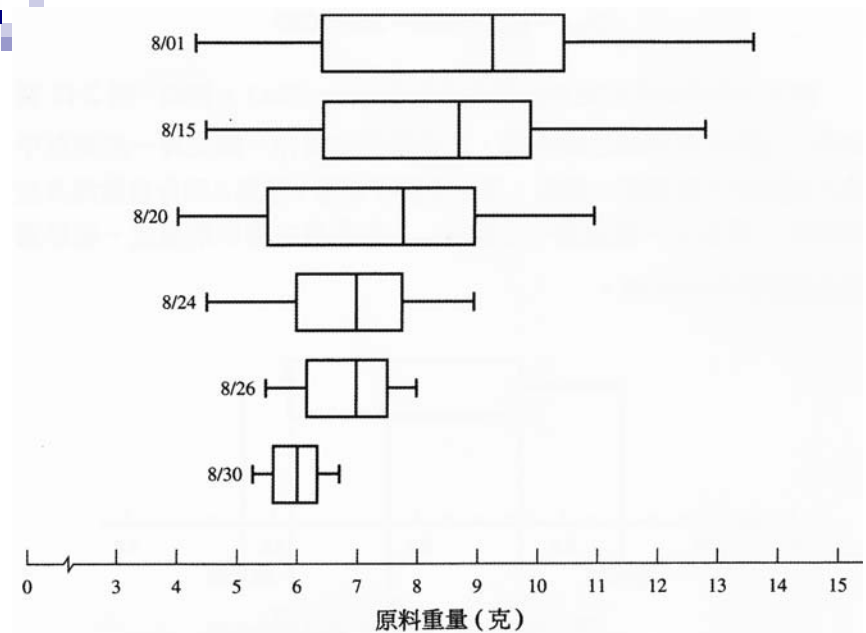


圖 2-17 由連續性衣料重量的樣本所繪製之箱形圖

組 界 (秒)	出現次數 f_k	組中點 X_k	X_k^2	$f_k X_k$	$f_k X_k^2$
11.0-小於 12.0	2	11.5	132.25	23.0	264.50
12.0-小於 13.0	16	12.5	156.25	200.0	2,500.00
13.0-小於 14.0	29	13.5	182.25	391.5	5,285.25
14.0-小於 15.0	27	14.5	210.25	391.5	5,676.75
15.0-小於 16.0	11	15.5	240.25	170.5	2,642.75
16.0-小於 17.0	6	16.5	272.25	99.0	1,633.50
17.0-小於 18.0	4	17.5	306.25	70.0	1,225.00
18.0-小於 19.0	2	18.5	342.25	37.0	684.50
19.0-小於 20.0	2	19.5	380.25	39.0	760.50
20.0-小於 21.0	1	20.5	420.25	20.5	420.25
總和				1,442.0	21,093.00

$$\bar{X} = \frac{\sum f_k X_k}{n} = \frac{1,442.0}{100} = 14.42 \text{ 秒}$$

$$s^2 = \frac{\sum f_k X_k^2 - n\bar{X}^2}{n-1} = \frac{21,093.00 - (100)(14.42)^2}{100-1} = 3.0238 \text{ 秒}^2$$

■ 母體變異數 (Population Variance) ，以希臘字母 σ^2 表示

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

其中 X_i 表示為第 i 個觀測值，分母 N 為母體大小 (SAT 例子中， $N=5$)。

■ 樣本變異數 (Sample Variance) ，以 s^2 表示

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{n-1}$$

由於母體平均數 μ 通常為未知，而以樣本的 \bar{X} 來推估，因此上式中分母以 $n-1$ 來除 (考慮失去一個自由度)。

表 2-8 變異係數與偏態係數之計算表列

任務項目	平均數 (0.01分鐘) \bar{X}	中位數 (0.01分鐘) m	標準差 (0.01分鐘) s	變異係數 v	偏斜係數 SK
1	11.139	9.833	3.338	0.346	1.174
2	5.604	4.613	2.354	0.420	1.263
3	2.540	1.908	0.588	0.232	3.224
4	4.229	3.133	1.068	0.253	3.079
5	9.957	9.081	2.141	0.215	1.227
6	2.913	2.068	1.665	0.572	1.523
7	2.576	1.858	1.451	0.563	1.484
8	5.990	5.070	2.021	0.337	1.366
9	4.467	3.295	2.435	0.545	1.444
10	2.969	2.432	0.881	0.297	1.829
11	3.532	2.790	1.039	0.294	2.142
12	4.465	3.698	1.446	0.324	1.591
13	5.903	4.736	1.832	0.310	1.908
14	3.305	3.197	1.087	0.329	0.298
15	3.210	2.520	1.446	0.452	1.432
16	5.984	5.362	1.789	0.299	1.043
17	4.126	3.378	1.678	0.219	1.337
18	7.270	6.461	2.034	0.263	1.193
19	2.770	2.133	1.063	0.384	1.797
20	6.673	5.914	1.968	0.295	1.157
21	8.530	7.833	1.845	0.216	1.052
22	3.928	3.325	1.688	0.430	1.072
23	5.247	4.604	1.469	0.280	1.313
24	5.094	4.402	1.234	0.242	1.682
25	5.452	4.750	2.079	0.381	1.013
26	3.653	2.958	0.982	0.269	2.123

水利工程之應用範例

2.1 Histograms (長條圖)

一組觀測或實驗數據，可用 histogram 或 frequency diagram 表示製作程序：

1. 收集觀測或實驗數據 Table 1-1
2. 最大值 (67.72) 至最少值 (39.91)，取整數 70~38) 與適當間距 (4 in)。間距數目 k 之取得決定於樣本數 n ，Iman and Conover(1983) 建議：

從 $2^k > n$ k 為最小整數

3. 計算每一個間距發生次數，並計算其發生頻率如表 1-2

4. 將變量劃於橫軸，發生次數劃於縱軸即為 Fig 1-1a

Fig 1-1b：發生次數之百分比 (發生次數除於總次數)

Fig 1-1c：frequency diagram

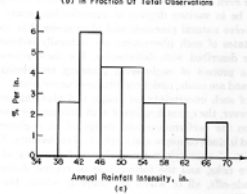
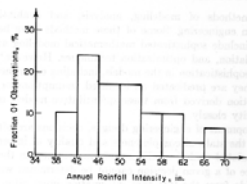
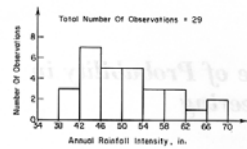
Table 1.1.

Year	Rainfall intensity (in.)
1918	43.30
1919	53.02
1920	63.52
1921	45.93
1922	48.26
1923	50.51
1924	49.57
1925	43.93
1926	46.77
1927	59.12
1928	54.49
1929	47.38
1930	40.78
1931	45.05
1932	50.37
1933	54.91
1934	51.25
1935	39.91
1936	53.29
1937	67.59
1938	58.71
1939	42.96
1940	55.77
1941	41.31
1942	58.83
1943	48.21
1944	44.67
1945	67.72
1946	43.11

ROLE OF PROBABILITY IN ENGINEERING

Table 1.2.

Interval	Number of observations	Fraction of total observations
38-42	3	0.1034
42-46	7	0.2415
46-50	5	0.1724
50-54	5	0.1724
54-58	3	0.1034
58-62	3	0.1034
62-66	1	0.0345
66-70	2	0.0690
Total = 29		1.0000



2.2 Quantile plots

1. Quantile plot portray the quartiles, or percentiles of the distribution of sample data. Quartile plots have following advantages :

- (1) All of the data are displayed
- (2) Every point has a distinct position

2. Construction

To construct a quartile plot

- (1) The data are ranked from smallest to largest
- (2) The smallest data value is assigned a rank $i=1$, while the largest receives a rank n
- (3) Each data is given a plotting position

Commonly-used formulas are :

$$\text{Weibull} \quad \frac{i}{n+1}$$

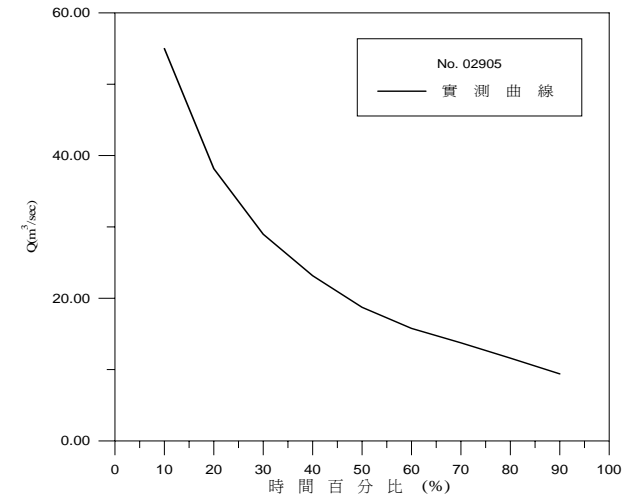
$$\text{Hazen} \quad \frac{i-0.5}{n}$$

$$\text{Gringorten} \quad \frac{i-0.44}{n} + 0.12$$

(4) The data values themselves are plotted along horizontal axis.
The plotting position of data is plotted on the other axis

(5) example : Flow duration curve (流量延時曲線)

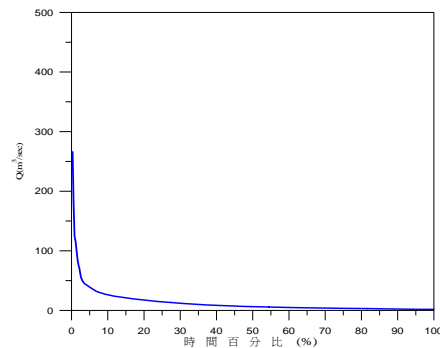
1. 定義：為一累積機率曲線，表示在某一特定期間內，高於或大於某一流量之百分比時間。(提供河川全年流量變化工具)



2. 製作：

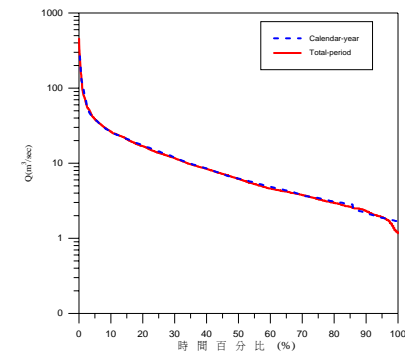
(a) the calendar-year method (範例)

Rank	Q _{year1} (cms)	Q _{year2} (cms)	...	Q _{year10} (cms)	Q _{ave} (cms)	P(%)
1	362.38	453.75	...	81.44	266.05	1/365 × 100% = 0.28%
2	239.58	79.78	...	54.17	181.55	2/365 × 100% = 0.56%
...
...
...
...
365	1.75	1.59	...	3.07	1.67	365/365 × 100% = 100%



(b) the total-period method(範例)

Rank	Q (cms)	P (%)
1	453.75	1 / 3650 × 100 % = 0.028 %
2	385.00	2 / 3650 × 100 % = 0.056 %
...
...
...
...
3650	1.17	3650 / 3650 × 100 % = 100 %





(a)(b) 兩者差別為 calendar-year method 在高流量部分會偏低，但在低流量部分會偏高。