

第二章 敘述統計(I)圖表法

本章內容

2-1 資料的型態

2-2 衡量的尺度

2-3 次數分配表

2-4 時間序列圖

數據是統計學應用的核心，從數據中萃取出有用的資訊是統計學的主要目的。因此，踏入統計學這門學們，當然必須面對真實的數據，雖然我們所處理的數據的量，相對於整體研究對象是少量的，然而如果沒有稍加整理，也很難得到有意義的資訊。例如，有一個半導體廠日產一千片 12 吋晶圓，

表 2-1 是抽測 50 片晶圓氧化層(oxide layer)晶圓厚度資料(um)：

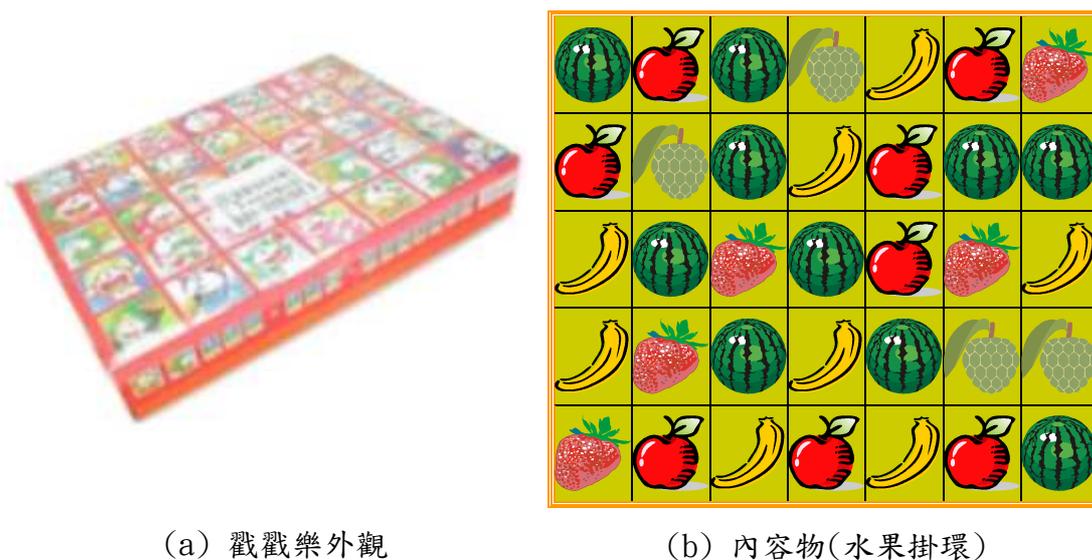
表 2- 1：氧化層晶圓厚度

1516	1512	1472	1449	1533	1505	1501	1500	1455	1554	1448	1462	1482	1412	1500
1568	1470	1509	1460	1483	1428	1427	1425	1517	1460	1512	1543	1488	1513	1385
1601	1470	1459	1451	1569	1484	1548	1520	1474	1566	1483	1547	1379	1461	1469
1495	1465	1614	1546	1461										

50 片僅是每日產量的 20 分 1，但

表 2- 1 的數據資料沒有經過有意義的整理，我們很難從這些生產數據，對整體製程有概略的描述與瞭解。所以，數據整理是統計科學應用的第一步。前段我們提到，整理必須適有意義的呈現數據的結構。要瞭解何為有意義的整理，就必須回到統計學根本的目的在於了解研究對象的全體在這量測變數上的分布

(distribution)。例如，圖 2- 1 (a)是一個小朋友過年常玩的戳戳樂遊戲盒的外觀，圖 2- 1 (b)為每一個格子內裝的一個水果掛環，共有 5 種，當然我們不會真的知道哪個位置放置哪一種水果掛環，圖 2- 1 (b) 這只是一種假想而已。



(a) 戳戳樂外觀

(b) 內容物(水果掛環)

圖 2- 1：兒時戳戳樂

在國小一年級的教材裡，要學生計算每一種水果出現的次數，我們可以得到表 2- 2，說明戳戳樂盒中每一種水果吊環的個數，稱為次數分配表 (frequency table)。

表 2- 1 讓我更清楚每一種水果吊環被戳中的機會，它代表每一種出象(outcome)的分布，它協助我們從混亂的圖 2- 1 中，對戳戳樂的內容物有更清楚且全面的瞭解。當然，戳戳樂這個遊戲，每一種水果吊環的次數分配，會隨著遊戲的進行而一直改變。

表 2- 2：戳戳樂各種水果掛環個數

吊飾種類					
個數	10	8	8	5	4

表 2- 1 和圖 2- 1 都是統計資料，沒有經過適當整理很難得到有意義的資料。所以，本章將介紹一些統計資料的整理方法，就如前所述，我們必須掌握統計資料整理的原則在於呈現資料的分配。而整理方法會隨著資料型態的不同而有差異或限制，因此，在介紹這些方法之前，我們先認識統計資料的分類方式。

2-1. 資料的型態

大學新生入學時，學校會給同學一張基本資料表填寫；或者，我們到醫院初診時，在掛號櫃台上通常也需填一填病患資料卡(表)，如表 2- 3 所列。表 2- 3 的每一個問題，我們稱為變數 (variable)，病患在變數上的填答內容稱為變量。通常在設計表格的每一個欄位(變數)時，就事先預期受調查者(病患)間在此欄位的填答內容應不盡相同，而一個有意義、必須的問題，應該預期有兩個(含)以上的變量。譬如，性別有男生和女生之別，但若是一間女子學校要入學學生填寫性別欄，則有些多餘。

表 2- 3：醫院初診病患資料表

姓名		生日		性別	
血型		身高		體重	
體溫		血壓(高)		血壓(低)	
居住地				聯絡電話	
學歷		職業		婚姻狀況	

表 2- 3 共有 14 個變數，依據病患可能填答變量的性質，可區分為量的資料與質的資料。以下我們說明如何分辨兩種資料屬性：

- 質的資料 (qualitative data)：『質』就是性質的意思，所謂性質就是與他人不同的特性，目的作為區分，分類。故又稱為類別資料 (categorical data)。表 2- 3 中，姓名，性別，血型，居住地，學歷，職業，學歷，和婚姻狀態這些變數所得到的資料，都屬此類。
 - 血型有 O, A, B 和 AB 型之別。
 - 婚姻狀況分成已婚與未婚之別。
 - 學歷有小學，國中，高中，大學，研究所以上之別。
- 其他常遇到的類別變數有宗教信仰，種族膚色，政治傾向等等。

● 量的資料：『量』有數量的意思，衡量個體對一個變數擁有的多寡、大小。表 2-3 中，身高、體重、體溫、血壓等。生日變數是很特別的一個變數，它可以依據研究者的目的，轉換成屬『質』或屬『量』的資料。星座愛好者會將生日轉換成 12 星座之一，然後分析星座與個性的關係。醫生通常將生日轉換成年齡，此時生日就成為屬『量』的資料。所以，我們必須注意變數的性質並非完全絕對，有時必須考慮到如何運用的問題。又如居住地，它可以是區別病患的來源，但它也可以反應出醫院服務的範圍，因為居住地可以轉換成距離，距離不是僅限平面上的直線距離，病患到醫院的交通距離可能更適切。屬量資料又可分為離散型(discrete)和連續型(continuous)資料。

■ **離散型資料**。通常是指計數的資料，例如本學期班上人數，本周缺曠人數，不良品個數，某日車禍發生次數，死亡人數，新生兒人數等等。這類變數的名稱通常已經明指出它是計數變數。

■ **連續型資料**。連續者，接續不斷之意，是一個極為抽象的概念。量測實體具有連續不斷之意，量測者意欲量測到最精確的程度，此一資料可稱為連續。我們無法由任何一組資料的呈現，確實表現出連續的意義。數字有『小數點』是一個有別於離散型計數變數的資料呈現，因為計數是一個一個累加，必不會有小數。但是我們仍然無法從任何一組數據中，看出連續不斷樣態。

■ **重要觀念**。

我們不應從資料出現表徵，作為辨識變數是否連續的依據，而它往往只是一種測量上的限制。譬如，身高 175 公分和 175.3 公分，我們不能因為前者有沒有小數點，就說身高是一個離散型變數，後者有小數一位就說身高一連續型變數。身高是一個長度單位，紀錄的有效位數是量測系統精確度的問題，不是變數本質的問題。長度本質是連續的，所以身高就是連續變數。同理，體重，溫度，時間，血壓等變數都是連續變數。考試分數是一個很奇怪的變數，有時它是一分一分往上累加，但有時考問答題或計算題，老師的給分會出現小數，但反應出老師考核學生學習

成效衡量的精確度。所以，若是考題是選擇題或填充題，考試分數對應答對的題數，它是一個離散變數。最小單位不一定是一分，例如老師考 50 題的選擇題，每題 2 分，則學生成績的最小單位就是兩分。若有問答題，則考試得分可以看作連續變數。

如上述關於變數的討論，我們將變數的分類架構繪於圖 2- 2。

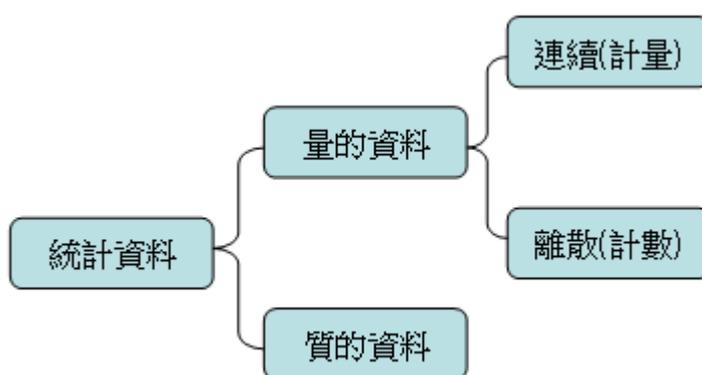


圖 2- 2：統計資料的型態

2-2. 衡量的尺度

前一節我們依據變數的特性分成屬性與屬量資料。在屬量資料上我們也探討到隨著研究者的主題，變數有可能是屬量或是屬性。本節我們將從另一種角度來討論變量的分類，反映對變數衡量的精確度，或是解析的程度來討論，統稱為衡量的尺度 (scale)。

譬如，身高變數是一個連續型的變數，但它可以有不同的衡量方式，比較粗略的，也是日常生活談話中常出現的，如果某人說李宜生是高個子和矮個子就是身高得一種衡量。但是這樣的說法，除非大家的經驗相同，我們很難從個人對高或矮的陳述，產生具體的想像。不過，類似的陳述卻是日常生活中極為常見的，譬如在民意調查中，受訪者被要求對問題回答是否滿意，或者在有限的等第區分下給予評分，可是每個人心中衡量的尺卻無從比較。

接續前例，當我們可以說李宜生身高超過 175 公分，這樣子就產生了比較具體的概念了。可是比起李宜生身高為 180 公分的陳述，前兩種關於身高的描述都不夠精細。所以，身高這個變數是一個連續型的變數，這個性質不與變量的紀錄

方式有關，但變量的紀錄方式卻表現出橫量的精確度。可是對於一個變數，我們並不一定能選擇採用何種尺度來衡量，有時還是受限於變數本身描述具體內容。例如，我們可以用粗略的尺度衡量身高或是體重，卻不太能用比男女變量更為精細來衡量性別。或許非正式言論中有“這個人很 man，或是這個人很娘”，但非客觀的衡量尺度。

一般而言，變量的尺度分成名目尺度 (nomial scale)，順序尺度 (ordinal scale)，區間尺度 (interval scale) 與比例尺度 (ratio scale)。尺度的精細程度依序為

名目尺度 → 順序尺度 → 區間尺度 → 比例尺度

所謂精細的程度就是可以將群體中每一個個體分別不同的程度。較精細的尺度可以降為較粗略的尺度，反之則不可。

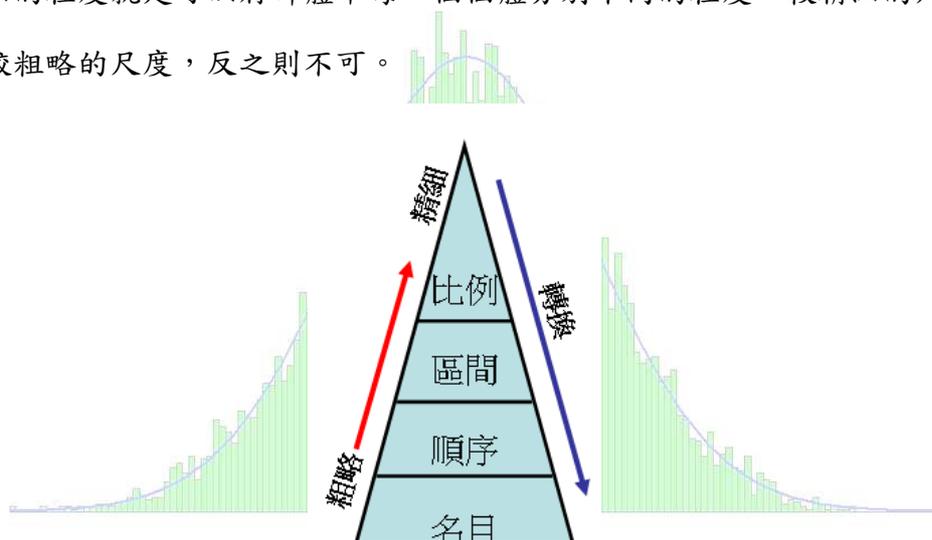


圖 2- 3：尺度精細程度與可轉換方向

- **名目尺度**：此變量尺度旨在區分不同。這是人類知識發展最初，最簡單的模式，辨識類屬。例如，性別有男和女之分。就是一種群體的分類，差異性的辨識。又如，宗教信仰有基督，天主，佛教，回教，道教等等類屬。膚色有棕，黑，白和紅等區別。此種只為將群體分成數個子群體的衡量方式，即稱為名目尺度。(如圖 2- 3)
- **順序尺度**：這是名目尺度更微細緻的尺度，除了分辨不同外，還進一步排除程度上的差異。常用的順序尺度有教育程度，國小、國中，...

等程度上的差異，又如問卷調查常問的滿意度，也是一種順序尺度。通常順序尺度據有名目尺度的用途，但往往名目尺度無法轉換成順序尺度，若為此，可能會造成很大的災難。例如，將種族的分類轉換成一種順序尺度，就可能造成種族歧視，種族屠殺等問題。然而，這類問題在人類史上幾乎無法避免，同學必須小心任何缺乏公平正義性質，僅以人種區分的政治主張。

- **區間尺度：**這裡引入變化量的觀念，除了順序尺度的程度差異，更把差異的程度數量化為大小。數量相差的大小相同，其意義或規模就相同。溫度通常使用區間尺度來衡量， 36°C 和 35°C 相差 1°C ， 106°C 和 105°C 也相差 1°C ，兩者意義相同。這點與順序尺度不同，高中學歷與國中學歷相差一個級數，大學學歷與高中學歷也差一個級數，但兩者沒有共同的意義存在。區間尺度相對的差異性可以進行比較。除了溫度外，出生的西元年也是一種區間尺度。張三 1968 年生，王五 1986 年生，兩人相差 20 歲。
- **比例尺度：**這是一個統計資料中最精細的尺度，數字的大小就有絕對資訊，無須採用相對差異性進行比較。例如，身高 200 公分的長人為身高 100 公分的矮人的兩倍，兩倍只兩個矮人堆疊一起與長人同高。比例尺度與區間尺度都是量化資料，它們間的差別為區間尺度為相對差異性，只要保有相對差異性的意義即可，所以，變量的零點也只是一個相對參考座標，是人定的可以依需要、習慣而改變。例如，溫度衡量的方式台灣與美國就不用的零點，但不影響討論每日氣溫變化的情形。比例尺度描述絕對性的差異，零點就是代表無，空無一物。長度為零，體重為零都是無的概念。但是溫度為零並非沒有溫度，它只是一個人造的參考原點。另外，以西元年紀錄生日，是利用耶穌生日作參考點。

表 2- 4：常用變數的衡量尺度

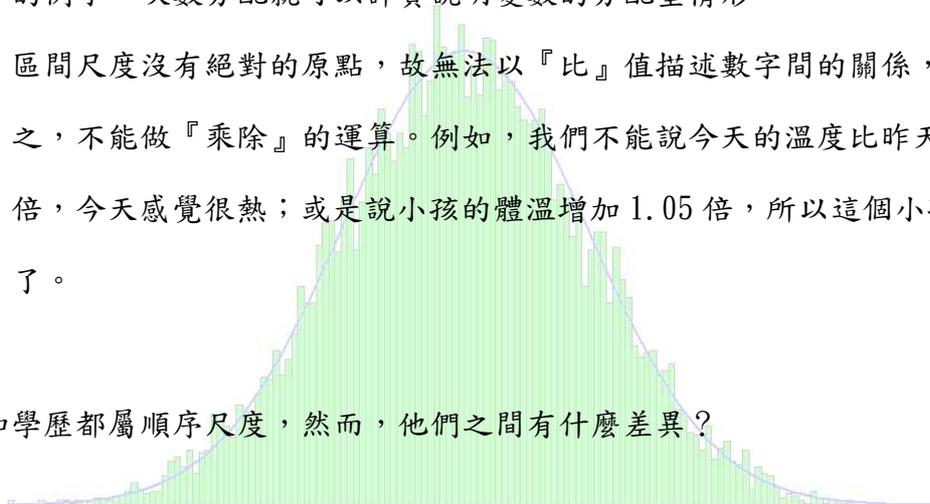
衡量尺度	常見例子
名目尺度	姓名、血型、宗教信仰、膚色、婚姻狀況、職業
順序尺度	滿意度、教育程度、等級
區間尺度	溫度、西元年
比例尺度	身高、體重、時間、壓力、次數、個數、人數

衡量尺度與四則運算的關係

- 名目和順序尺度沒有量的意義，故在資料處理上，不能使用加減乘除四則運算。通常合適的做法，僅計算每一個類別觀察到的次數，如表 2-2 的例子，次數分配就可以詳實說明變數的分配型情形。
- 區間尺度沒有絕對的原點，故無法以『比』值描述數字間的關係，換言之，不能做『乘除』的運算。例如，我們不能說今天的溫度比昨天高兩倍，今天感覺很熱；或是說小孩的體溫增加 1.05 倍，所以這個小孩發燒了。

討論：

滿意度和學歷都屬順序尺度，然而，他們之間有什麼差異？



2-3. 次數分配表

前一節我們介紹了統計變數的性質與變量衡量的尺度，這節我們將依據不同的資料型態學習如何整理我們蒐集到的數據。這裡還是再次強調，整理數據的目的在於呈現資料分布的型態，這裡介紹幾種方法，但不限定只有這些方法，特別是各行各業處理的數據都不一樣，所以，資料整理呈現也會些差異，然而，都不能違反整理數據的原始目的。最基本的方式也是最直接的方式就是以圖或表格來描述一筆數據。

屬性資料的次數分配。如前一節戳戳樂的例子，計算每一種掛環出現的次數，以表格呈現就是次數分配表。一個屬性資料的次數分配表，包含兩個主要部分：

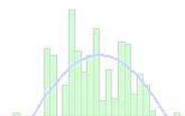
每一種出象(outcome)的次數和合計次數。將次數除以合計次數，所得之數值定義為相對次數 (relative frequency)。

$$\text{相對次數} = \frac{\text{次數}}{\text{合計次數}}$$

表 2- 5：掛環種類的次數與相對次數表

吊飾種類						合計
次數	10	8	8	5	4	35
相對次數	0.29	0.23	0.23	0.14	0.11	1.0

例 2- 1



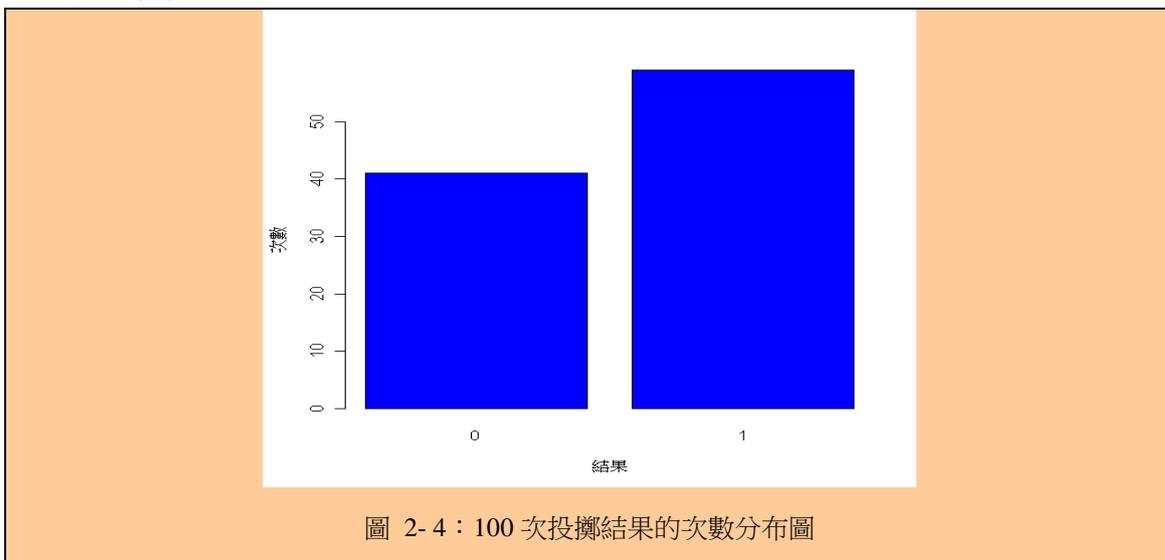
銅板出現正面紀錄為 1，出現反面紀錄為 0。以下是投擲一個五元銅板 100 次的結果：

01111111100101001101000101011100010011111001101110

11111101111101001100101110000011100000111101101011

結果	次數	相對次數
正面(=1)	59	0.59
反面(=0)	41	0.41
合計	100	1.00

利用劃記法可以得上表，並利用 R 軟體繪製下圖。我們稱下圖長條圖(bar chart)。



R 程式語法：

```
> x
[1] 0 1 1 1 1 1 1 1 1 0 0 1 0 1 0 0 1 1 0 1 0 0 0 1 0 1 0 1 1 1 0 0 0
1 0 0 1
[38] 1 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 1 0 0 1 1 0 0
1 0 1 1
[75] 1 0 0 0 0 0 1 1 1 0 0 0 0 0 1 1 1 1 0 1 1 0 1 0 1 1
> y<-table(x)
> barplot(y,col="blue",border=T,ylab="次數",xlab="結果")
```

例 2-2



2008 針對本校學生的問卷調查，384 受訪學生中的血型分布如下表。

血型	次數	相對次數
A	80	20.8
B	74	19.3
AB	48	12.5
O	176	45.8
未知	6	1.6
合計	384	100.0

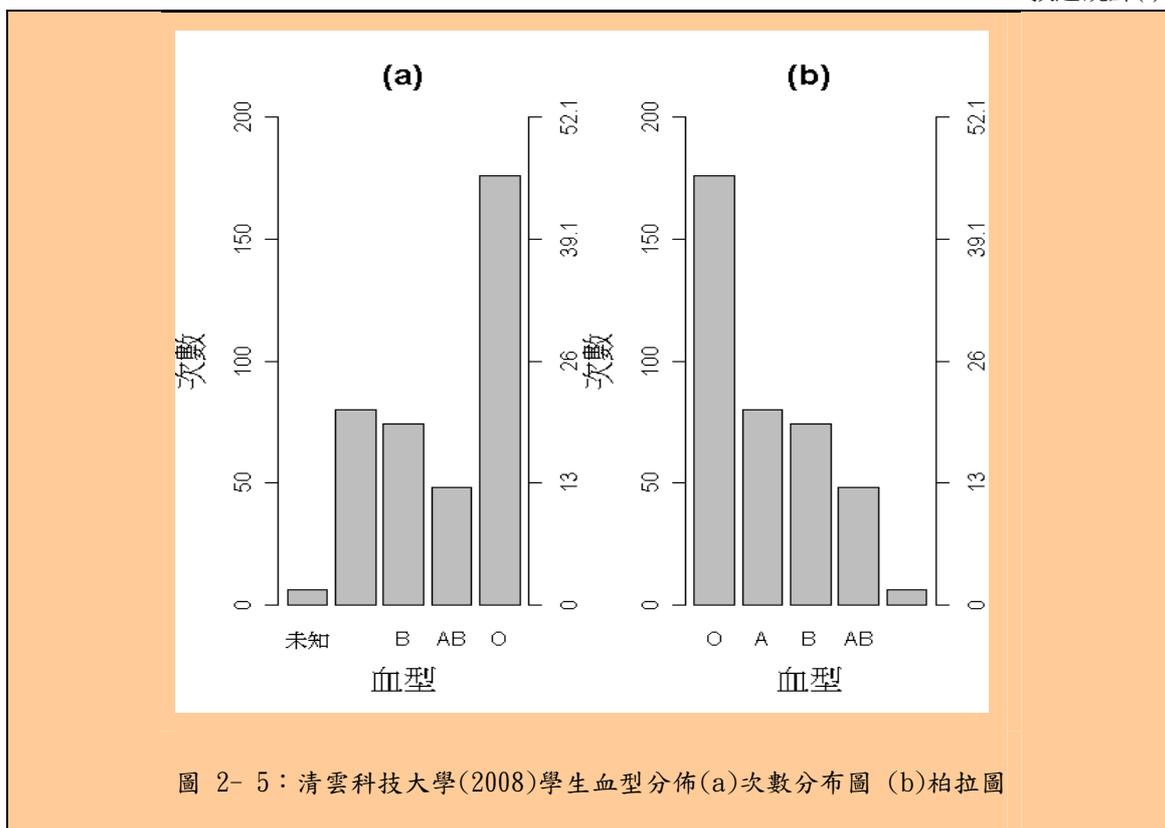
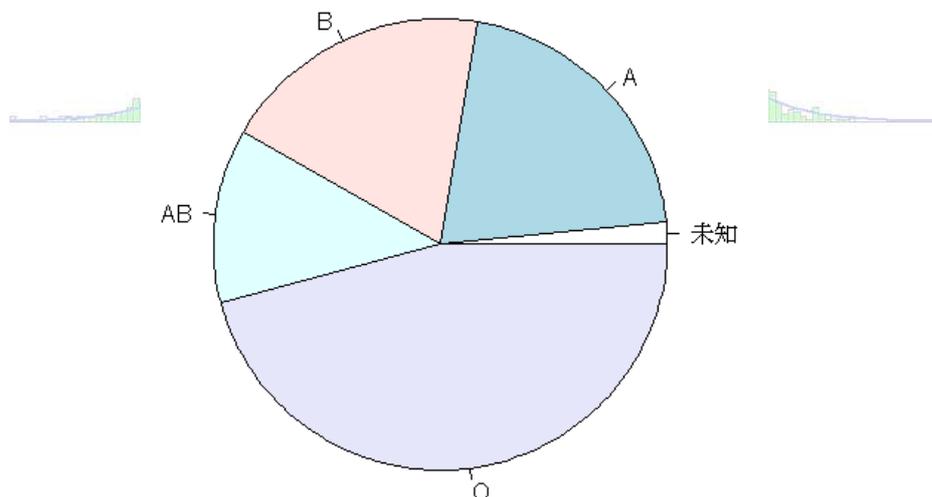


圖 2- 5：清雲科技大學(2008)學生血型分佈(a)次數分布圖 (b)柏拉圖

註：柏拉圖是一種長條圖，依據圖形長條排列順序依據出現次數多寡由左至右排列。當選項多時，柏拉圖可協助查覺比例較高的幾個類屬。



2- 6： 血型分布的圓形圖

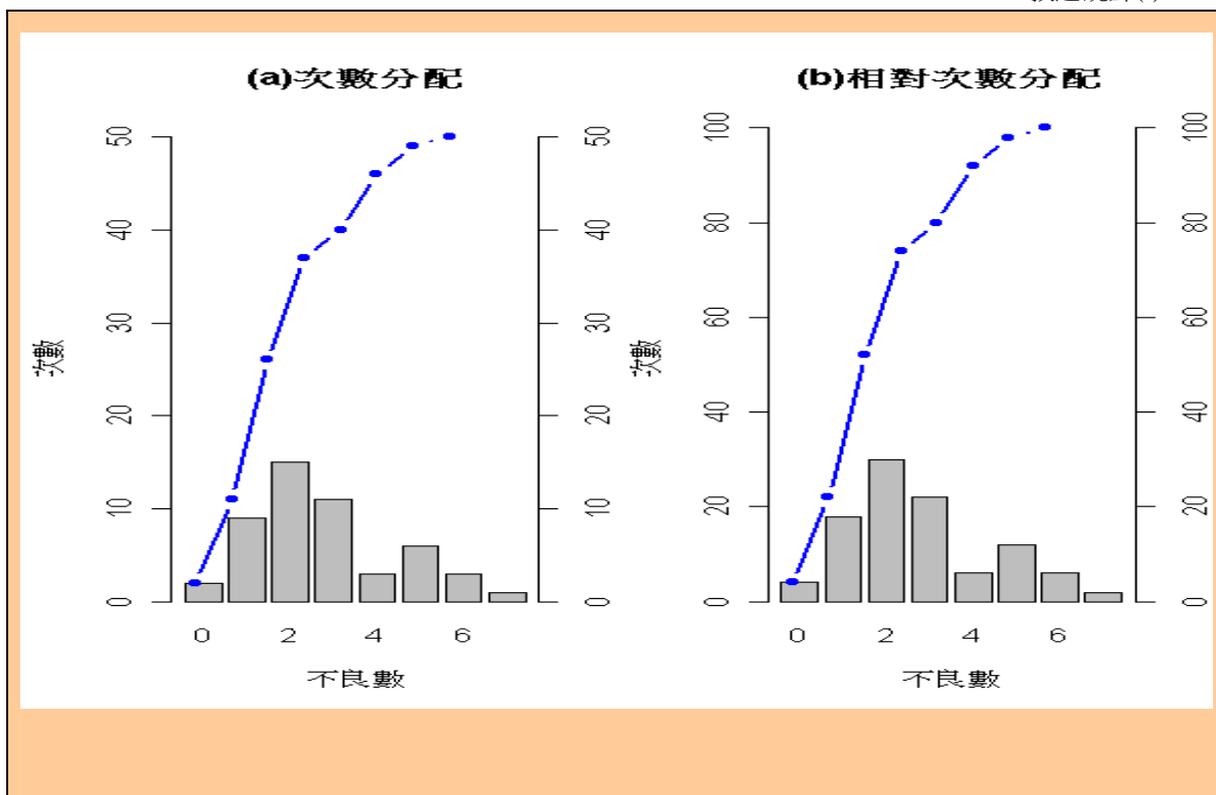
除了長條圖外，圖 2- 6 的形圖(pie chart)更能呈現每一個類屬在全體中所佔的比例

某工廠檢測 50 批產品，每批不良品數量紀錄如下

6 1 1 6 2 1 2 2 2 2 7 2 0 2 3 3 1 2 5 0 1 3 1 1 3
 2 2 3 4 4 3 5 2 5 5 1 4 3 3 5 5 6 2 3 3 3 2 2 1 2

不良數	畫記	次數	累積次數
0	//	2	2
1	### ////	9	11
2	### ### ###	15	26
3	### ### /	11	37
4	///	3	40
5	### /	6	46
6	///	3	49
7	/	1	50
合計		50	

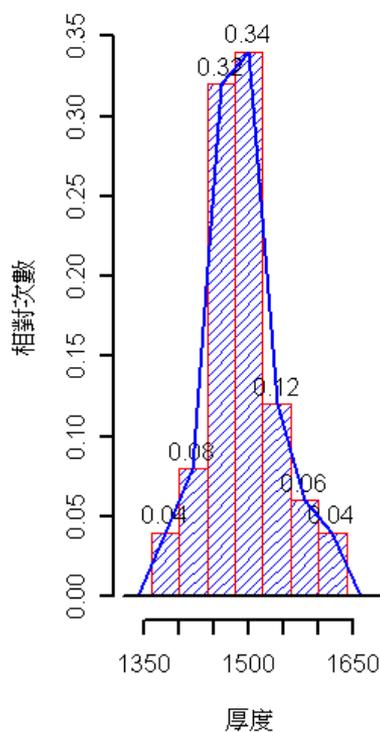
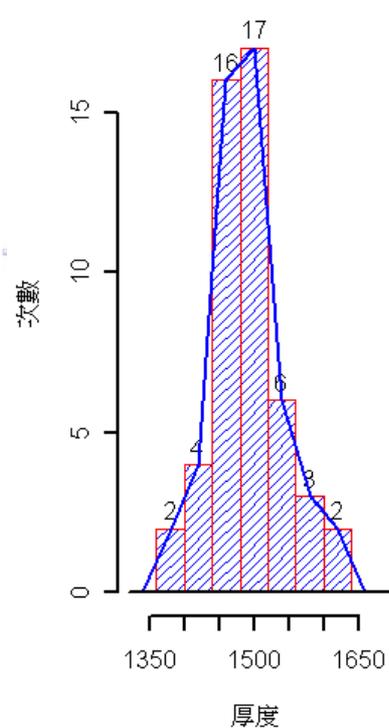
不良數	次數	累積次數	相對次數 (%)	累積相對次數 (%)
0	2	2	4	4
1	9	11	18	22
2	15	26	30	52
3	11	37	22	74
4	3	40	6	80
5	6	46	12	92
6	3	49	6	98
7	1	50	2	100
合計	50		100	



連續資料分組的步驟：

- 1 從數據中，找出最大數據 (Max) 與最小數據 (Min) 來。
- 2 計算數據的全距 $R = Max - Min$ 。
- 5 決定適當組數 k (5~20)，與組距 h 。注意： $k \times h > R$ 。
- 6 求出最小組的下組界 $s_1 < \min$ 。注意 $s_1 < \min$ ，即最小組的下組界必須小於或等於 Min ，
- 7 算出每一組之下組界 s_i 與上組界 l_i ， $i = 1, \dots, k$ 。
- 8 將數據畫記於表內。

組別	範圍	組中點	次數	累積 次數	相對 次數(%)	累積相對 次數(%)
1	1360~1400	1380	2	2	4	4
2	1400~1440	1420	4	6	8	12
3	1440~1480	1460	16	22	32	44
4	1480~1520	1500	17	39	34	78
5	1520~1560	1540	6	45	12	90
6	1560~1600	1580	3	48	6	96
7	1600~1640	1620	2	50	4	100
合計			50		100	



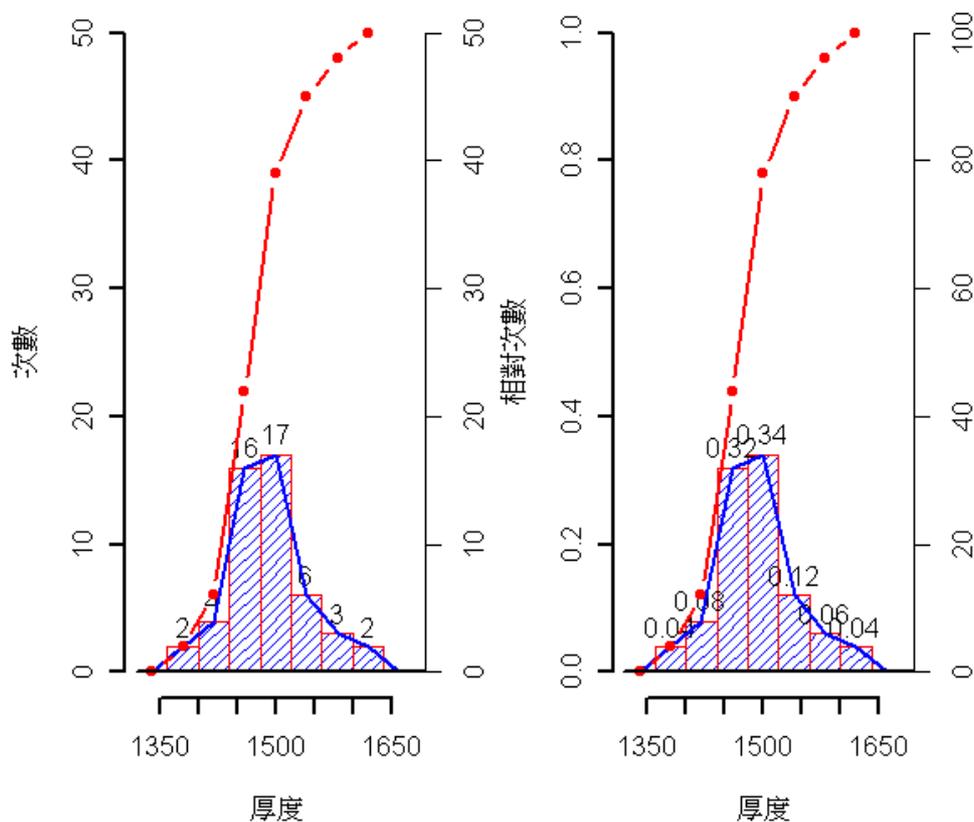


圖 2- 7：晶圓厚度的次數分配圖

R 程式語法：

```
br<-0:7*40+1360 # 設定次數分配的組界分隔點
hist(x, breaks=br, xlab="厚度", ylab="次數", main="", col="blue", angle=45, density=20,
lwd=2, labels=TRUE, border="red", cex.lab=1, ylim=c(0,18))
```

R 程式語法：#多邊形圖語法

```
y<-hist(x, breaks=br, xlab="厚度", ylab="次數", main="", col="blue", angle=45,
density=20, lwd=2, labels=TRUE, border="red", cex.lab=1,
ylim=c(0,18), xlim=c(1320,1680))
lines(c(1340, y$mids[1]), c(0, y$counts[1]), lwd=2, col=" blue" )
for(i in 1:6){
lines(c(y$mids[i], y$mids[i+1]), c(y$counts[i], y$counts[i+1]), lwd=2, col=" blue" )
}
lines(c(y$mids[7], 1660), c(y$counts[7], 0), lwd=2, col=" blue" )
lines(c(1320, 1700), c(0, 0), lwd=2, col=" black" )
```

2-4 時間序列圖

統計資料的另一種分類方式是以時間做為區分，考慮資料是否與蒐集時間有關聯性。前幾節討論的資料，都不考慮資料蒐集過程的先後順序，稱為橫斷面資料。倘若研究的問題是特別注意時間趨勢，也就是某種現象隨著時間變動情形，在資料蒐集時特別記載發生的時間點，通常稱此種資料為時間序列。

近年來台灣人口成長趨緩的議題受到教育界普遍的關心，影響各級學校未來生存發展，圖 2-8 為 1981 年以來我國新生兒人數的趨勢圖，2008 年新生兒人數已降至約 20 萬，僅有 1981 年約 40 萬的一半。

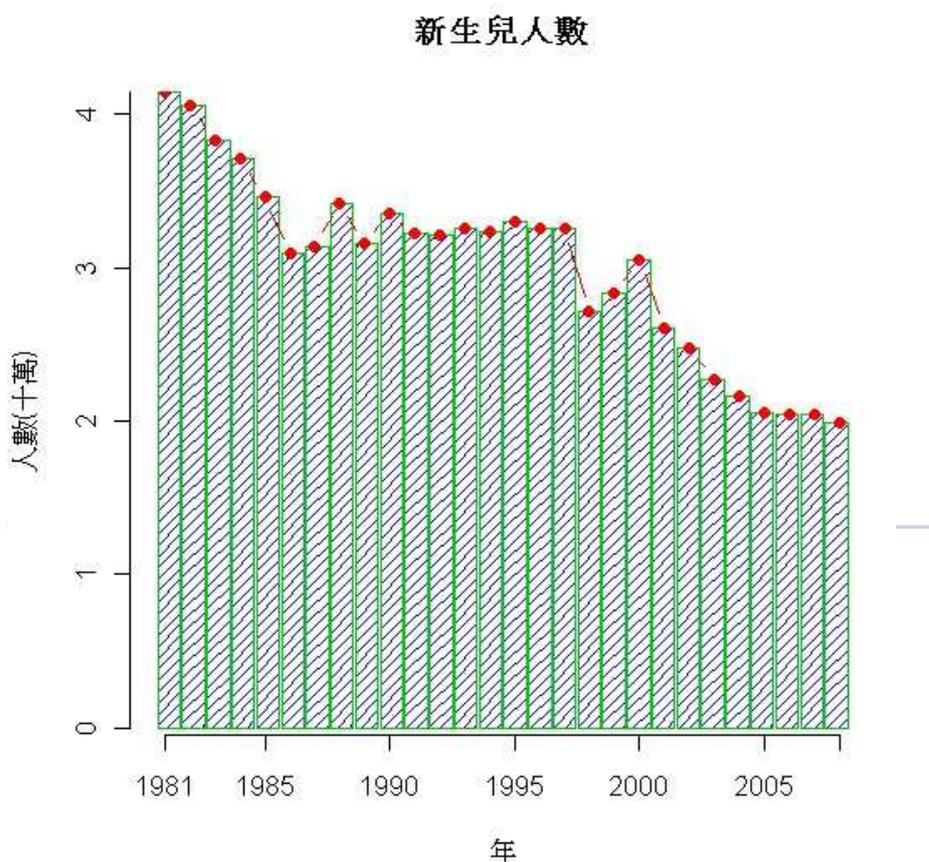


圖 2-8：1981 年以來我國新生兒數的趨勢圖

隨著去年(2008)由美國引發的全球金融海嘯(financial downturn)另一個令人注目的議題就是失業率的變化。圖 2-9 是我國 2008/7 以來失業率的變化趨勢圖，顯示有逐漸遞增的趨勢。

失業率

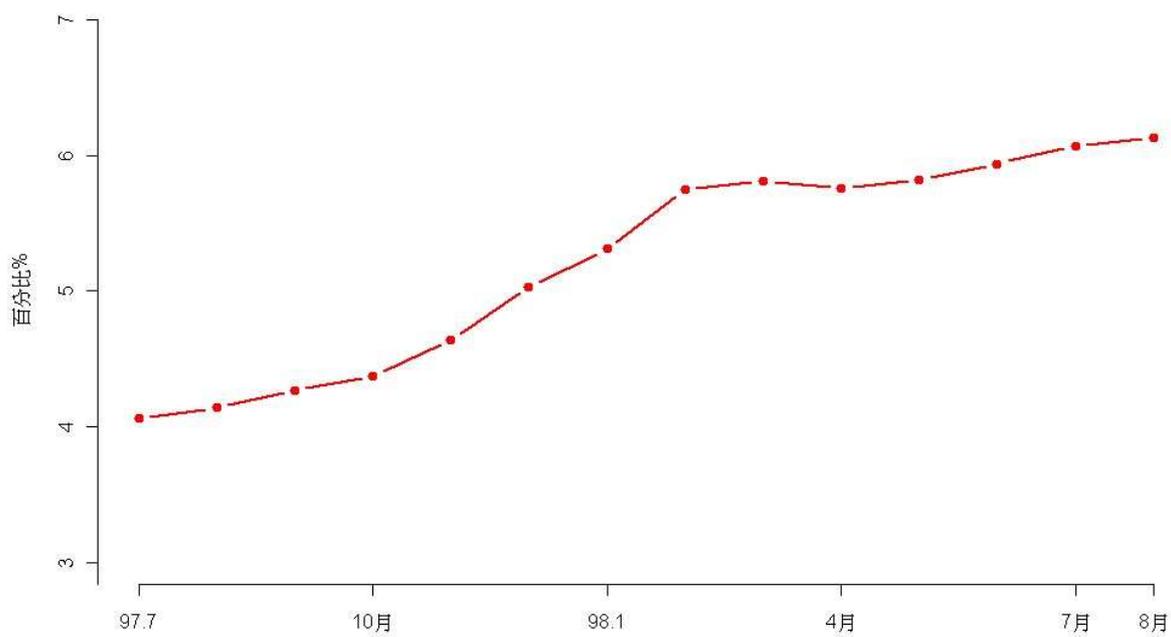


圖 2-9：2008/7 以來的失業率變化

參考資料

行政院主計處 www.dgbas.gov.tw



習題

- 下列何者屬於離散型數據：(A) 厚度 (B) 尺寸不良數 (C) 寬度 (D) 壽命 (E) 身高。
- 量測鋼板厚度是屬於何種數據：(A) 離散 (B) 連續型數據 (C) A、B 皆是 (D) 計數值 (E) 間斷。
- 量測後，統計不符合規格之鋼板個數屬於何種數據：(A) 計數值 (B) 間斷型數據 (C) A、B 皆是 (D) 計量值 (E) 連續型數據。
- 請用一個例子說明區間尺度與比例尺度的差異。
- 請用一個例子說明名目尺度與順序尺度的差異。
- 請指出下列變數的屬性(量或質)？

(a) 每日學生缺課人數。

(b) 每日 H1N1 新增人數。

(c) 出生嬰兒的體重。

(d) 病患的血型。

(e) 學校社團的類別。

- 請繪製下列資料的次數和累積相對次數分配表、直方圖與累積相對次數分配圖。

46.4 66.9 65.1 46.7 64.9 62.0 62.2 61.8 60.2 48.8

80.0 56.5 52.7 71.4 42.2 79.0 55.0 47.2 57.0 51.9

55.0 61.6 48.5 54.9 61.2 63.7 61.6 65.5 61.9 42.8

67.2 50.1 49.6 52.8 39.3 60.7 53.8 68.7 52.9 66.2

54.9 61.4 57.7 56.4 57.3 48.8 59.2 59.9 64.3 64.2

- 調查本校 40 位同學每日的上網時間(小時)，資料如下。繪製次數和相對次數分配表、直方圖與累積相對次數分配圖。

0.9 1.0 1.0 1.0 1.0 1.1 1.2 1.2 1.3 1.4

1.4 1.4 1.4 1.4 1.5 1.5 1.6 1.8 1.8 1.8

1.9 2.0 2.0 2.0 2.0 2.0 2.1 2.1 2.3 2.4

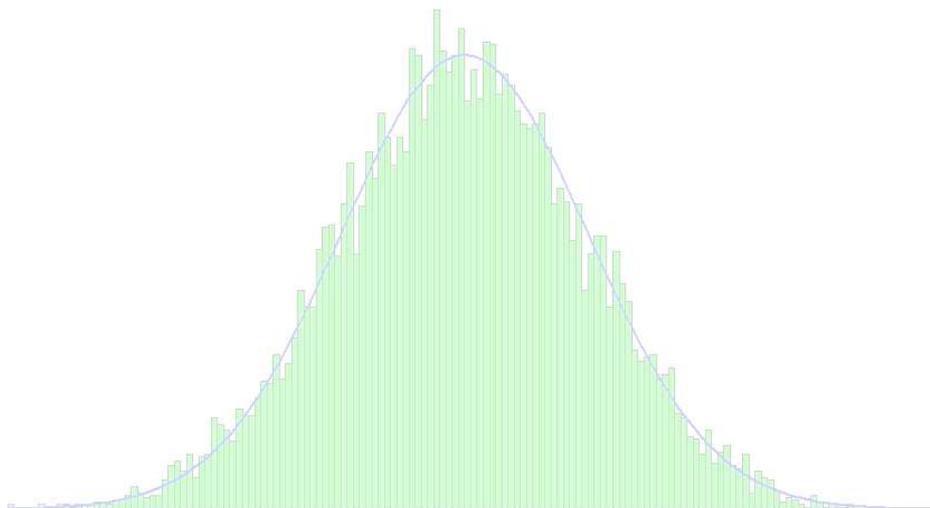
2.5 2.6 2.6 2.7 2.7 2.7 2.8 3.1 3.3 3.4

6. 下表係本校某問卷調查各學院學院和各年級的受訪人數。

學院	人數	年級	人數
管理	112	一	100
工	87	二	98
商	86	三	108
電資	99	四	78
合計	384		384

(a) 請計算各學院受訪學生的相對次數分佈表，繪製長條圖與圓形圖。

(b) 請計算各年級受訪學生的相對次數分佈表，繪製長條圖與圓形圖。



重點摘錄

