第 4 章　檢視資料的分布

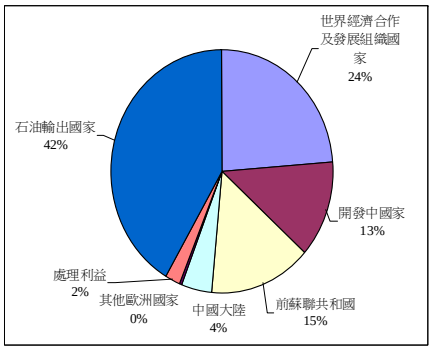
以統計表與統計圖呈現

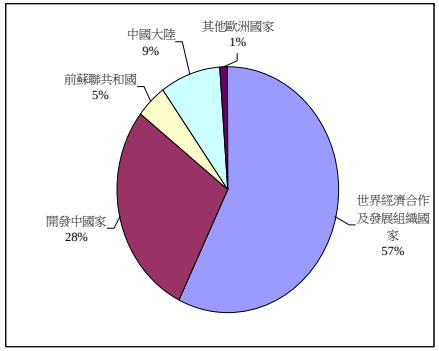
- 若資料只有少數幾個觀察值(數據)，一眼就可看出資料的特質。然而，大部分資料的觀察值是大量的、片段的、複雜的、未分類或分組的，此時就需藉助統計學的方法，將觀察值分類或分組，編製**統計表**或**統計圖**，以便觀察資料的特性。

資料整理的方法——次數分配

- **次數分配**(frequency distribution)：將觀察值分類、分組，並計算觀察值在各類或各組出現之次數的統計方法，稱為次數分配。
- 次數分配是將類別資料中的觀察值依類別分類並計算各類別的次數，或將數量資料中的觀察值依大小做有系統的排列，然後分組並計算各組的次數。經過分類或分組後，可以顯示觀察值在各組的分布情形。底下即是一例：

- 例子：底下分別是 2007 年石油的生產(左)與消費(右)百分比圓餅圖，由圖中可清楚看出世界各經濟體的石油生產及消費狀況。OPEC 生產最多的石油，OECD (已開發國家)消費最多石油。





類別資料的整理與資料呈現

- 類別資料的次數分配：依照類別分別排列，並計算各個類別的次數的統計表稱為類別資料的次數分配表。

- 類別資料在整理時首先可依類別分組，計算各組的次數，然後以統計表顯示資料在各組的分布情形。
- 分組時應兼顧**互斥**(觀察值只能屬於其中一組，不能屬於兩組或以上)與**周延**(觀察值必須屬於其中一組)兩個原則。
- 觀察值在分組之後通常會失去一些資訊，所以分組時應特別注意，不要因為分組而喪失有用的資訊。
- 例子：青少年對媒體的偏好。假設我們隨機抽樣 48 個青少年，調查其最喜歡的媒體，資料結果如下。如何整理這些資料呢？

網際網路	手機	手機	報紙	報紙	報紙	有線電視	有線電視
有線電視	有線電視	網際網路	網際網路	手機	網際網路	報紙	報紙
雜誌	有線電視	雜誌	有線電視	有線電視	手機	有線電視	手機
網際網路	手機	網際網路	雜誌	雜誌	報紙	網際網路	雜誌
有線電視	網際網路	手機	手機	報紙	網際網路	有線電視	手機
網際網路	有線電視	網際網路	網際網路	手機	網際網路	報紙	雜誌

統計表

- 統計表：將蒐集得到的資料整理成表格的形式，並以文字或數字的形式表現出來，即是所謂的統計表。
  - 統計表可以有系統、有條理的方法，表現出資料的主要內容及特性，讓讀者一目了然，提供有意義與有用的資訊。
- 一個完整的統計表至少應包括：
  - 標題(title)：包括表號(表序)與標題。
  - 標目(label)：標目是用來表示表身所要表示的項目或事實。
  - 表身(body)：表身是資料的主體，是統計表的核心。
  - 資料來源及附註：應標明資料來源出處，以方便讀者查閱。

類別資料的次數分配表

- 類別資料的**次數分配表**(frequency table)：依照類別分別排列，並計算各個類別的元素出現的次數的統計表稱為類別資料的次數分配表。

- 例子：青少年媒體偏好的粗資料(或未分組資料)可分類整理成底下的**次數分配表**，由該表可立即看出青少年對各種媒體的偏好順序：網際網路、有線電視、手機...。

變數→

媒體類別	人數 ←次數欄
網際網路	13
有線電視	11
手機	10
報紙	8
雜誌	6
合計	48

資料來源：自行抽樣調查

類別資料的相對次數

- **相對次數**(relative frequency)：各類別的次數比例。

$$\text{某類別的相對次數} = \frac{\text{某類別的次數}}{\text{所有類別的次數總合(總次數)}}$$

- 例子：青少年媒體偏好的**相對次數分配表**，可立即得知喜歡網際網路的比例最高，其次為有線電視。相對次數分配表比(絕對)次數分配表容易表達變數的特性，更易看出資料的全貌。

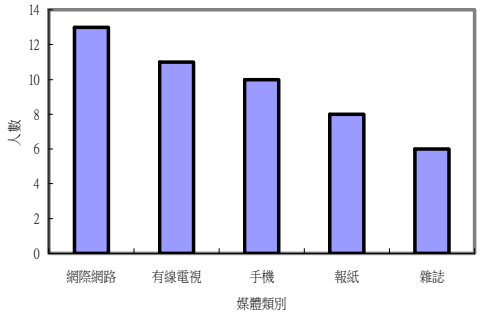
媒體類別	相對次數	百分比%
網際網路	0.27	27
有線電視	0.23	23
手機	0.21	21
報紙	0.17	17
雜誌	0.12	12
合計	1.00	100

類別資料的統計圖

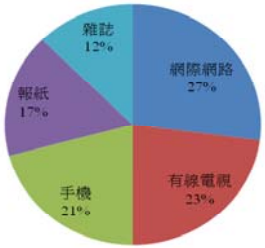
- 統計圖：將資料以點、線、面、體等圖形為主，以文字數字為輔的表現方式即為統計圖。意即利用點的多寡，線的長短粗細、起伏趨勢，面積與體積的大小，顏色深淺來表示資料的特性者稱之為統計圖。

  - 統計表雖可用來呈現資料的分布，但統計圖可以將分布情況更清楚地表達出來(人類辨識圖形的能力優於文字)，亦更容易找出圖形的一般型態(shape)，找出有異於一般型態的離群值(outlier)。
- 統計圖的種類：
  - **線圖**(line chart)：描繪變數在不同的時點所衡量出來的結果，以直線的高低來表示資料的特質，最常用來表示時間序列資料的演化，亦稱時間序列圖(time series plot)。
  - **長條圖**(bar chart)：以長條的長短來表現資料的特性，通常用於類別資料，又稱煙囪圖。
  - **圓餅圖**(pie chart)：以一個像一塊餅的圓形來表示全部的資料，各部分資料則以佔整個圓餅的百分比來表示，類別資料和順序資料最常用圓餅圖來表示。
  - **直方圖**(histogram)：次數分配的長方形圖，又稱次數直方圖

  - 柏拉圖(Pareto chart)：以次數與累加百分比來表示類別資料特性的圖形。
  - 散佈圖(scatter diagram)：用來表示兩個或多個變數間的關係
  - 次數多邊圖：連結直方圖各組的組中點所形成的圖形。
- 類別資料的圖形：
  - **長條圖**：以長條的長短、高度或數值的大小來表示各個類別的次數的統計圖。



➢ **圓餅圖**：以整塊餅的圓形表示全部的資料，各部分表示各個類別的相對次數或百分比的統計圖稱為圓餅圖。



數量資料的整理與資料呈現

- 數量資料觀察值通常很多而且未分組。數量資料可依其數值的大小加以分組，將數值比較接近的歸為一組，先整理成次數分配表，再繪成次數分配圖，以觀察資料的分布情形。
- 數量資料的**次數分配表**：將數量資料分成若干個組，同時計算、列示各組次數的統計表稱為數量資料的次數分配表。

- 數量資料次數分配表的建立步驟

(1) 求全距：觀察值中的最大值減去最小值即為全距(range)

全距＝最大值－最小值

(2) 決定組數：決定將數值資料分成幾組。

- 組數的多寡視資料的範圍與特性而定；觀察值越多，組數應越多，一般分為 5~10 組。
- 組數的多寡是一個主觀的問題，並無客觀的標準。
- 若組數太多，則不容易看出分配型態及其變化趨勢，但可保持資料的真實性(若組數太多，則其圖型會呈現出扁平的直方圖)；若組數太少，則所得的次數分配表過於簡化，可能會失去次數分配的意義，無法顯示資料的特性(若組數太少，圖型會呈現出集中且高聳的直方圖)
- 若以 n 表示觀察值個數，則組數 k 通常要求滿足

$2^k \geq n$　　或是　$k \geq \log_2 n$

(3) 決定組距：組距是組與組間的距離，以及前後組之組下限與組上限之距離。

- 若組數已決定，可將全距除以組數得到一個約略的組距
- 組距一般採整數原則，最好是 2、5、10 的倍數，以方便計算。
- 各組組距最好相等。但若觀察值分佈的範圍很廣，大多數的觀察值集中於一個小範圍內，而其他少數觀察值分散在其他範圍內，此時組距可不相同；觀察值密集處其組距可縮小，觀察值分散處其組距可加大。
- 最好不要有開放組距(沒有下限值或上限值，僅以小於某上限值或大於某下限值來表示)，但第一組與最後一組可採開放組距

(4) 選擇上下限：分組時必須選擇各組的上限與下限，可由最小值開始，以組距建立組限，直到所有觀察值均包括在內為止

- 每一觀察值必須歸屬唯一的一組。
- 若採開放組距，第一組寫『上限以下』(20 以下)，最後一組寫『下限以上』(80 以上)。
- 連續資料的前後組上下限必須相等，以確保所有觀察值均分屬任何一組中；若觀察值等於前後組之上下限，應歸屬於後一組。連續資料的分組可表達為

組下限 $\leq x <$ 組上限

例如：$0 \leq x < 10$、$10 \leq x < 20$…

- 間斷變數(觀察值個數可數)前後組上下限不必相等，例如 $1 \leq x \leq 9$、$10 \leq x \leq 19$…

(5) 計算組中點：組中點(class midpoint)是各組上下限的平均數

$$\text{組中點} = \frac{\text{組下限} + \text{組上限}}{2}$$

(6) 計算各組次數：通常用劃計法(「正」字)逐一記錄觀察值屬於哪一組，最後計算各組的次數。

- 例子：海之濱冰店的營業收入。該冰店 4、5 兩個月的營業收入如下表(單位為百元)，觀察值個數有 61 個。冰店老闆想了解冰店的營收分布狀況，欲建立**次數分配表**。

	A	B	C	D	E	F	G	H	I	J	K
1	21	39	35	44	53	26	45	85	63	65	72
2	35	54	78	52	51	57	64	32	56	54	
3	36	53	74	37	42	25	42	34	83	59	
4	41	57	61	68	56	48	51	75	45	45	
5	55	45	57	77	48	54	67	62	55	55	
6	28	58	76	46	89	65	45	66	67	67	

  - 計算全距：最大值為 89，最小值為 21，因此全距為 89－21＝68。
  - 決定組數：因觀察值個數為 61，$\log_2 61 = 5.93$，滿足 $2^k \geq 61$ ($2^k \geq n$)的最小組數 k 為 6。
  - 決定組距：由於全距為 68，組數為 6，因此組距為 68÷6 = 11.33，約略為 10。故將組距設定為 10。

  - 選擇上下限：最小值為 21，因此第一組以 20 為組下限、以 30 為組上限，第二組之組限為 30 與 40，依此類推...，最後一組之組限設定為 80 與 90。
  - 計算組中點：組中點是上下限之平均，例如第一組之組中點為(20＋30)/2＝25。
  - 計算各組次數：使用『正』字劃計法紀錄觀察值屬於哪組

組號	組限	組距	組中點	劃記	次數
1	$20 \leq x < 30$	10	25	𠂇	4
2	$30 \leq x < 40$	10	35	正丅	7
3	$40 \leq x < 50$	10	45	正正丅	12
4	$50 \leq x < 60$	10	55	正正正𠂇	18
5	$60 \leq x < 70$	10	65	正正一	11
6	$70 \leq x < 80$	10	75	正一	6
7	$80 \leq x < 90$	10	85	𠂇	3
					$\sum f_i = 61$

相對次數分配

- 相對次數分配是指各組次數佔總次數的比例，亦即各組次數相對於總次數的比例分配。
- **相對次數**(relative frequency)之定義

$$\text{相對次數} = \frac{\text{組次數}}{\text{總次數}}$$

數學符號：若 f_i 表示第 i 組次數，n 為樣本觀察值的總個數，則第 i 組的相對次數 rf_i 定義為

$$rf_i = \frac{f_i}{n}$$

- 百分比分配：有時我們喜歡將比例表達成百分比的形式，例如 0.17＝17%；比例與百分比間的轉換公式如下

$$\text{百分比} = (\text{相對次數}) \cdot 100$$

- 例子：海之濱冰店營業收入的相對次數分配表。

組號	組限	次數	相對次數 rf	百分比%
1	$20 \leq x < 30$	4	4/61=0.07	7
2	$30 \leq x < 40$	7	7/61=0.11	11
3	$40 \leq x < 50$	12	12/61=0.20	20
4	$50 \leq x < 60$	18	18/61=0.29	29
5	$60 \leq x < 70$	11	11/61=0.18	18
6	$70 \leq x < 80$	6	6/61=0.10	10
7	$80 \leq x < 90$	3	3/61=0.05	5
			1.00	100

累加次數分配

- 當我們想知道某一水準以上或以下的次數總和時，可計算各組的累加次數。[例如想知道冰店營業額有多少次小於 4000 元]
- 以下累加次數(cumulative frequency; **CF**)：以下累加次數簡稱**累加次數**，以符號 CF_i 表示，指小於等於第 i 組的次數和。亦即

$$CF_i = f_1 + f_2 + \cdots + f_i$$

- 以上累加次數(decumulative frequency; **DF**)：以上累加次數是指大於等於第 i 組的次數和，以符號 DF_i 表示。亦即(假設組數等於 k 組)

$$DF_i = f_i + f_{i+1} + \cdots f_k$$

- 例子：海之濱冰店營業收入的累加次數分配。

組號	組限	次數	以下累加次數	以上累加次數
1	$20 \le x < 30$	4	4	3+6+11+18+12+7+4=61
2	$30 \le x < 40$	7	4+7=11	3+6+11+18+12+7=57
3	$40 \le x < 50$	12	4+7+12=23	3+6+11+18+12=50
4	$50 \le x < 60$	18	4+7+12+18=41	3+6+11+18=38
5	$60 \le x < 70$	11	4+7+12+18+11=52	3+6+11=20
6	$70 \le x < 80$	6	4+7+12+18+11+6=58	3+6=9
7	$80 \le x < 90$	3	4+7+12+18+11+6+3=61	3

累加相對次數分配

- 累加相對次數分配係指某一水準以上或以下的相對次數的總和。
- 以下累加相對次數(cumulative relative frequency; **CRF**)：以下累加相對次數，是指小於等於第 i 組的相對次數和。以符號 CRF_i 表示。亦即：

$$CRF_i = rf_1 + rf_2 + \cdots + rf_i$$

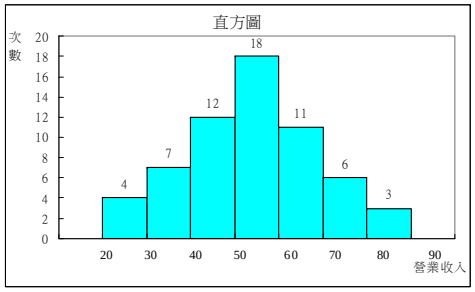
組號	組限	相對次數	以下累加相對次數	以上累加相對次數
1	$20 \le x < 30$	0.07	0.07	1.00
2	$30 \le x < 40$	0.11	0.18	0.93
3	$40 \le x < 50$	0.20	0.38	0.82
4	$50 \le x < 60$	0.29	0.67	0.62
5	$60 \le x < 70$	0.18	0.85	0.33
6	$70 \le x < 80$	0.10	0.95	0.15
7	$80 \le x < 90$	0.05	1.00	0.05
合計		1.00		

- 以上累加相對次數(decumulative relative frequency; **DRF**)：以上累加相對次數是指大於等於第 i 組之相對次數和。以符號 DRF_i 表示，亦即

$$DRF_i = rf_i + rf_{i+1} + \cdots + rf_k$$

數量資料的圖形呈現

- 數量資料的統計圖形可以呈現出資料的分布情況，如中心位置、對稱性、分散程度等。
- **直方圖**：表示次數分配的長方形圖，它是以 X 軸表示各組的組界，Y 軸為次數所畫出來的長方形圖，又稱為次數直方圖。
  - 絕對次數分配、相對次數分配均可以直方圖來呈現
- 例子：海之濱冰店營業收入的次數直方圖。(如何解釋直方圖？)
  - 觀察該圖的形狀(shape)可知，該直方圖有一個**尖峰**(peak)，出現在 50~60(百元)這一組；而且大致**對稱**(symmetric)，也就是說直方圖的分布型態在尖峰兩側很相似。

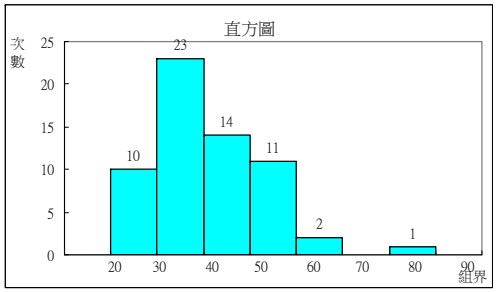


  - 每日營收的分布大致上是以 50~60(百元)為中心，然後向兩側分散。亦即，營業收入的中心點約在 55(百元)，距離此一中心點的低營業收入與高營業收入日數差不多，使得圖形左邊與右邊看起來差不多。
  - 離散程度：當資料很分散時，最大值與最小值與多數的觀察值有很大的差異時，我們說離差(分散程度)很大，此時不會出現尖峰，表示資料不集中。而當觀察值離散程度較小時，就會呈現出資料的一般型態(集中趨勢)。

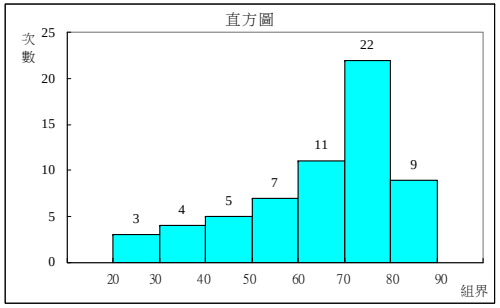
- 當資料離散程度很大時，一般而言，必須做進一步的處理[例如去掉**離群值**(outlier)：與大多數觀察值差異很大的觀察值]。
- 由以上直方圖可以看出，海之濱冰店的營業收入沒有離群值，資料呈現集中的型態

- **非對稱(asymmetric)分配**
  - 並非所有的資料分配情況均是對稱的。例如：之前的海之濱冰店日營收是屬於4、5月的春季時節，小額收入與大額收入的日子較少，因此分配呈對稱型態。但如果在炎熱的夏天或寒冷的冬天，營業受入的分配也許不再呈對稱分配。
  - 炎熱的夏天：吃冰的人應該較多，因此營業額相對較高的日子應該會比較多，營業額小的日子則相對較少。如下圖的直方圖所示，尖峰出現在70~80(百元)那組，大部分營收集中在該組附近，營業額比50(百元)小的日子並不多。

  - 圖形呈現出**右偏**(skewed to the right)的趨勢。
  - 右偏：直方圖右邊(相對較大的觀察值)延伸出去的部份比左邊遠得多[這是因為觀察值多集中在左邊]。
  - 圖中有一個**離群值**(outlier)[落在圖形的一般型態之外的觀察值]，該離群值使得圖形呈現非常明顯的右偏。
  - 當有離群值存在時，可將離群值撇開不看，或者將有離群值與沒有離群值兩者相互對照比較。

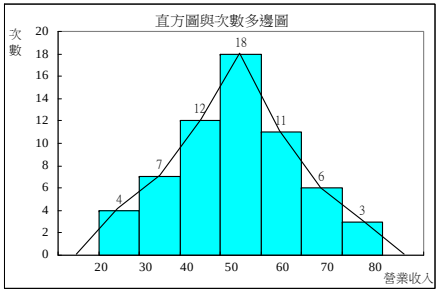


  - 圖形呈現出偏斜(skewed)的分布，而且是呈現出**左偏**(skewed to the left)的趨勢。
  - 左偏：直方圖左邊(相對較小的觀察值)延伸出去的部份比右邊遠得多[這是因為觀察值多集中在右邊]。

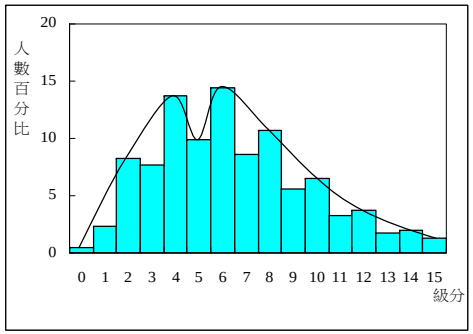


  - 寒冷的冬天：吃冰的人應該較少，因此營業額相對較小的日子應該比較多，營業額高的日子則相對較少。如下圖的直方圖所示，尖峰出現在30~40(百元)那組，大部分營收集中在該組附近，營業額比60(百元)大的日子並不多。

- **注意**：當我們觀察直方圖(或其他統計圖)時，主要是要找出尖峰的位置(資料集中的中心位置)與是否有離群值，看看是**對稱**還是具有**偏態**的。
- **次數多邊圖**：連結次數直方圖或相對次數直方圖各組的組中點，並前後各延伸半個組距單位即為次數多邊圖。
  - 次數多邊圖係根據直方圖而繪製，以線圖方式表示資料的分配型態。
  - 例子：海之濱冰店營業收入的次數多邊圖。[在連續變數的狀況下，通常以一條如次數多邊圖的線來代表資料的分布]



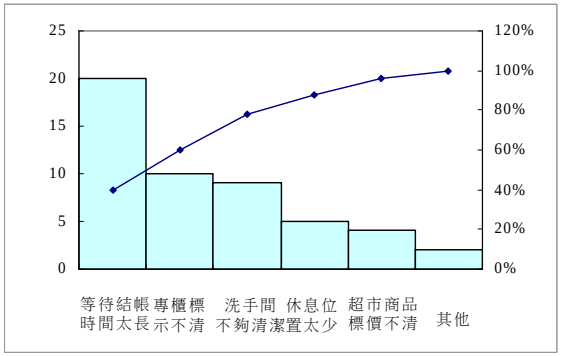
- **雙峰**(bimodal)：當資料有兩個尖峰出現時，稱為雙峰分配。
  - 例子：2007 年大學學測數學科成績的相對次數分配直方圖，共分 0~15 級分。圖中可看出兩個尖峰(4 與 6 級分)，此一現象顯示有兩個學生群的數學程度有明顯差異，該差異可能來自於居住區域(學校與城鄉差距)、家庭所得高低、或其他因素。



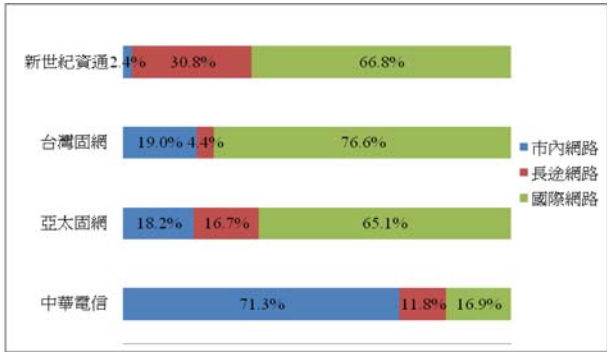
- **Pareto 分析**(柏拉圖)
  - 類別資料可以用柏拉圖來表示其特性。柏拉圖是以次數與累加百分比來表示類別資料特性的圖形，常用於品質管制。
  - 例子：某百貨公司的客服部想瞭解顧客對公司服務的意見，以作為改進的參考。在抽樣調查了 100 個顧客的意見以後，發現顧客抱怨的次數分配如下表所示。
    - 對品質管制的意涵：若想要降低顧客抱怨，最有效的方法是解決『等候結帳時間太長』與『專櫃標示不清』兩個問題。

	顧客抱怨的次數與百分比		
	次數	百分比（%）	累計百分比（%）
等待結帳時間太長	20	40	40
專櫃標示不清	10	20	60
洗手間不夠清潔	9	18	78
休息位置太少	5	10	88
超商標價不清楚	4	8	96
其他	2	4	100

  - 在畫柏拉圖(Pareto 圖)時，先將類別資料的次數依序由大至小排列，橫軸表示類別[依先前之次數大小排序]，左邊縱軸表示次數，右邊縱軸表示累加百分比，以直方圖(或長條圖)畫出。由柏拉圖可看出次數發生最多(最重要)的類別，亦可看出各類別的累加百分比，以了解各類別的重要性及累加類別的百分比。



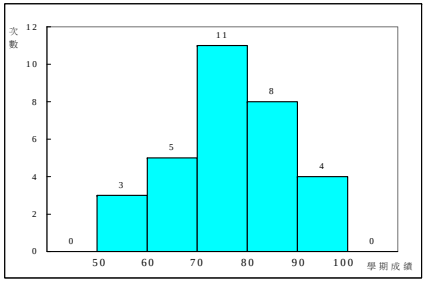
- **多段長條圖**：一次看幾個類別資料的相對次數分配。
  - 例子：台灣的固網業者有 4 家，中華電信、新世紀資通、亞太固網、台灣固網；固網業者的營業項目主要為室內網路、長途網路、國際網路。若想比較固網業者各項業務的營收比例，可使用多段長條圖：中華電信的固網營收以市話居多，其他三家業者則著重在國際電話。



- **枝葉圖**(stem and leaf plot)：將觀察值分成二部份，一部份為枝，另一部份為葉。枝的部份為高位數字，葉的部份為低位數字。
  - 若資料的數值為二位數，枝即是十位數(左邊位數)，葉是個位數(右邊數字)[通常葉只取一位數]，如數字 34 中，枝為 3，葉為 4。若數字太大(三位數以上)，有可能會造成枝幹太多，此時就需慎選枝幹個數[枝葉圖不好用之處]。
  - 例子：某班級計量經濟學的學期成績。中心位置：成績主要集中在 70~80 分；分散情形：最低成績 56 分，最高 95 分。

枝	葉
5	6 7 8
6	4 5 6 7 9
7	1 2 3 4 4 5 5 7 8 8 9
8	4 5 5 6 6 7 8 9
9	0 2 4 5

  - 若將枝葉圖旋轉 90 度，看起來就像直方圖一樣，所以枝葉圖可視為是橫躺著的直方圖[直方圖擁有的訊息並不會比枝葉圖少太多]。

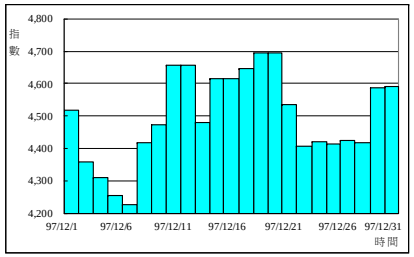


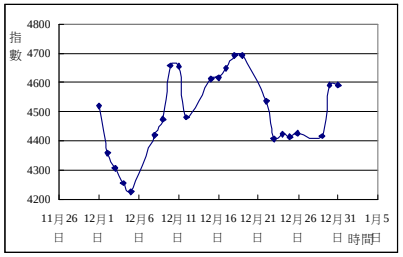
			9	
			8	
			8	
			7	9
			5	8
			5	7
		9	5	6
		7	4	6 5
	8	6	3	5 4
	7	5	2	5 2
	6	4	1	4 0
	5	6	7	8 9

  - 枝葉圖的優缺點：
    - 相對於次數分配直方圖而言，枝葉圖不會失去原有資料的資訊。
    - 枝葉圖僅適用於數量資料，且觀察值數目不多的情況。

繪製統計圖的注意事項

- 統計圖的主要目的：讓讀者一目了然，清楚的呈現出資料的分布，把數據當中重要的資訊顯示出來。
- 圖形的選擇：力求簡潔、清楚，呈現最重要的資訊。
- 例子：股價指數走勢。2008 年 12 月台灣加權股價指數 23 個交易日的資料，要用何種圖最能表達股價的變動趨勢？從長條圖可以看出每個交易日的指數大小，但線圖更能傳達股價指數的變化趨勢。





- 例子：平均國民所得的成長。若想要傳達民國 75~96 年間台灣平均國民所得快速成長的想法，最好是用右圖，因為它看起來比較陡峭，成長速度看來比左圖所呈現的要快。左右兩個圖只差在縱軸的長度(右圖的縱軸較長)，巧妙運用這種長度上的技巧即可創造出意想不到的效果。