統計，讓數字說話！

Statistics

concepts and controversies

Chapter 4

描述分佈

2009/07/04

分佈

分佈（Distribution）：

一個變數的分佈告訴我們：該變數的可能值有哪些，以及那些值發生的頻繁程度。

- 展示數據

  利用數據來傳達事實及支持結論，如同利用言詞一樣。除了很小量的資料外，所有資料都需綜合濃縮較精簡的形式，可以表或圖顯示。

- 展示分佈

  要展示用名目尺度度量的變數之分佈，可用圓辦圖或長條圖。

數據表

- **Ex. 怎麼樣的表才清楚？**

18歲以上婦女的婚姻狀況。

表4-1 成年婦女婚姻狀況，1994年

婚姻狀況	計數（單位千人）	百分比
單身	19,458	19.7
已婚	58,133	58.8
寡居	11,073	11.2
離婚	10,120	10.2
總計	98,765	100.0

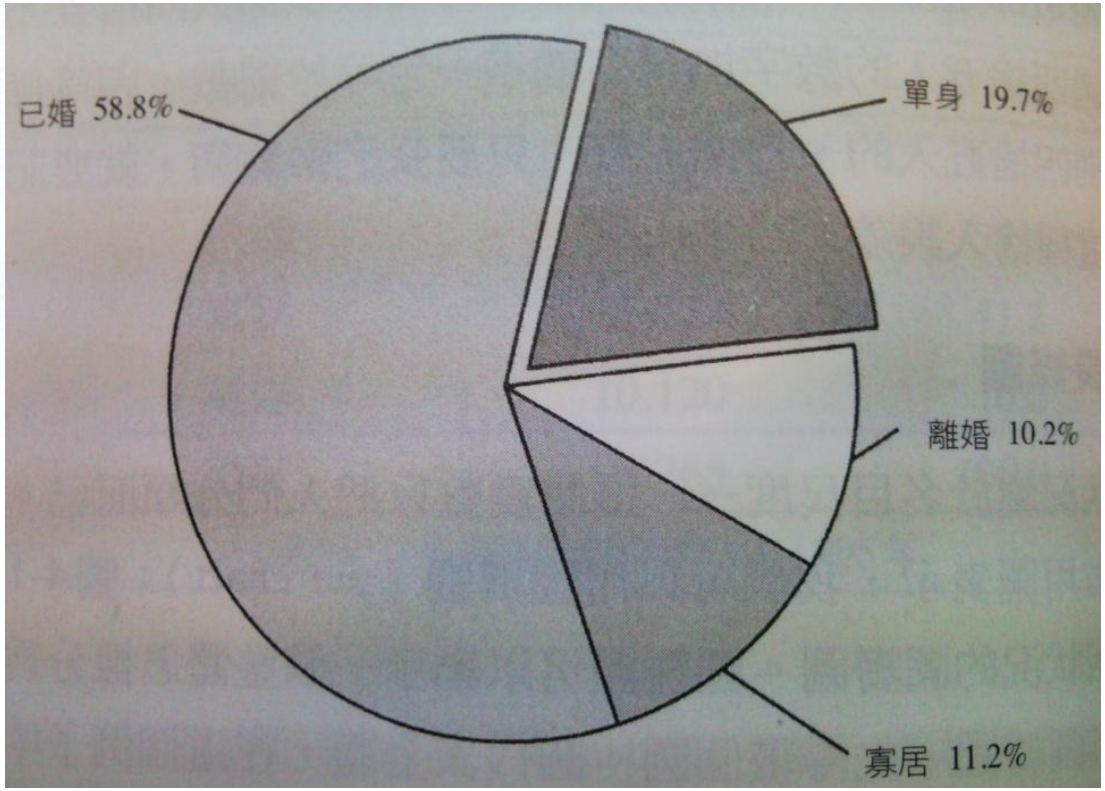
來源：《美國統計精粹》，1995年

→主題、年份、單位需明示。

圓瓣圖

婚姻狀況

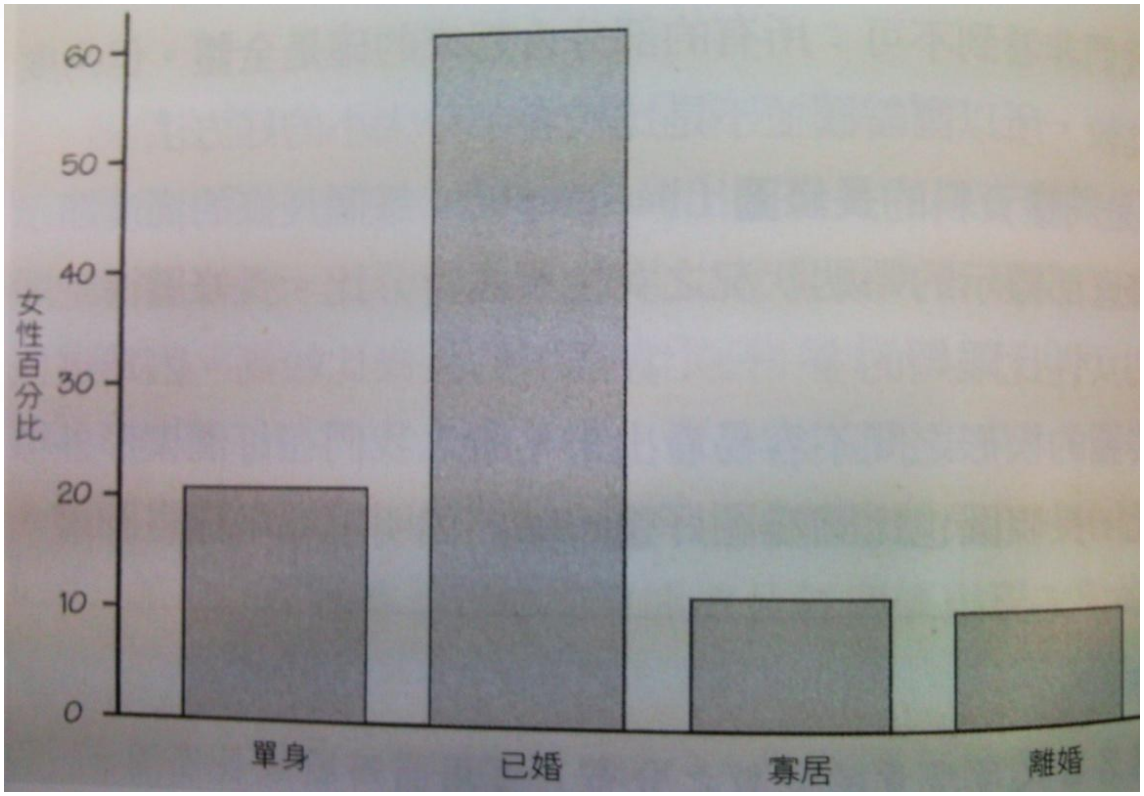
- 系屬名目尺度，該變數只作分類，可以圓瓣圖表示分佈
- 圓瓣圖強調各部份的計數或百分比與整體的關係。



長條圖

婚姻狀況

- 可以長條圖比較各部份大小，較圓瓣圖易顯示差異。
- 長條圖強調各部份彼此之間數量大小的比較。

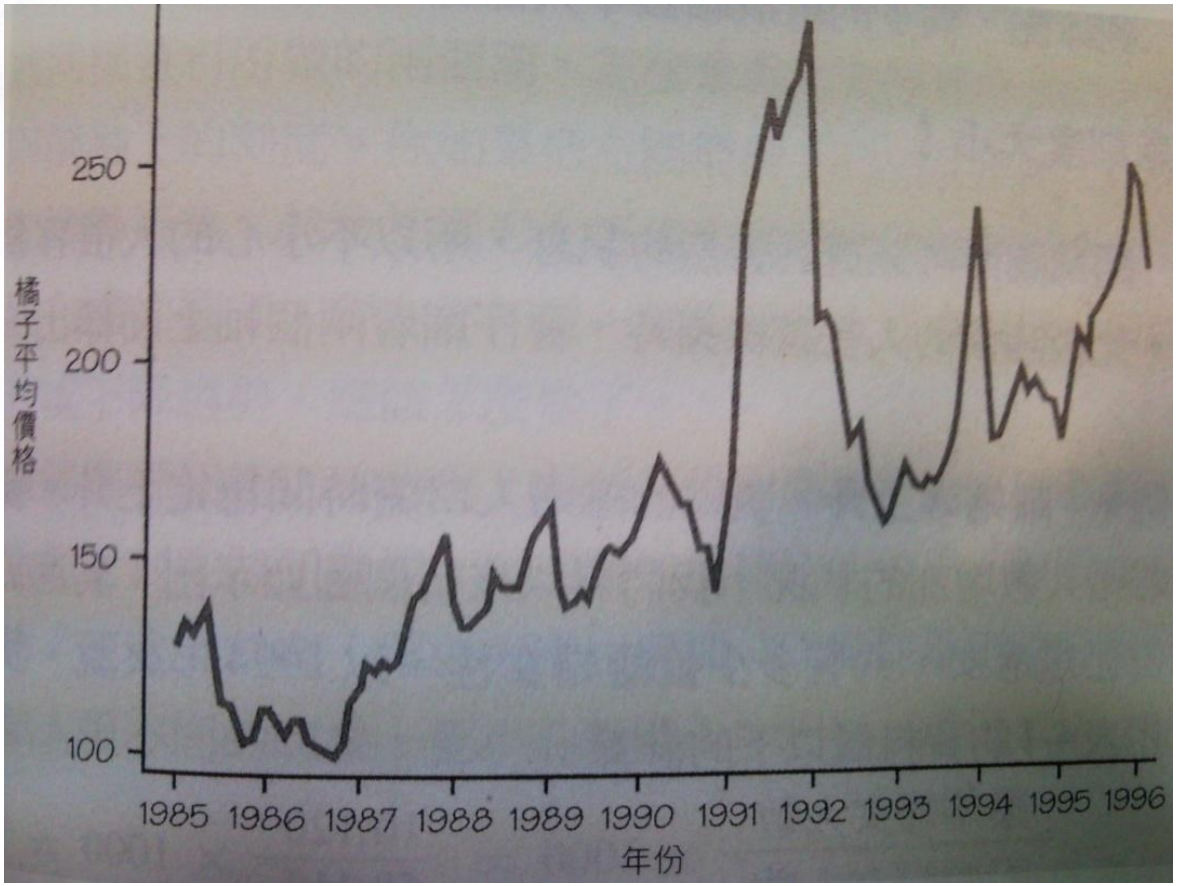


留意象形圖

- 當畫長條圖時，每個長條都需同寬。要是從藝術美感的觀點，長條圖實在有點單調。
- 象形圖其為長條圖，只是以圖形取代長條。

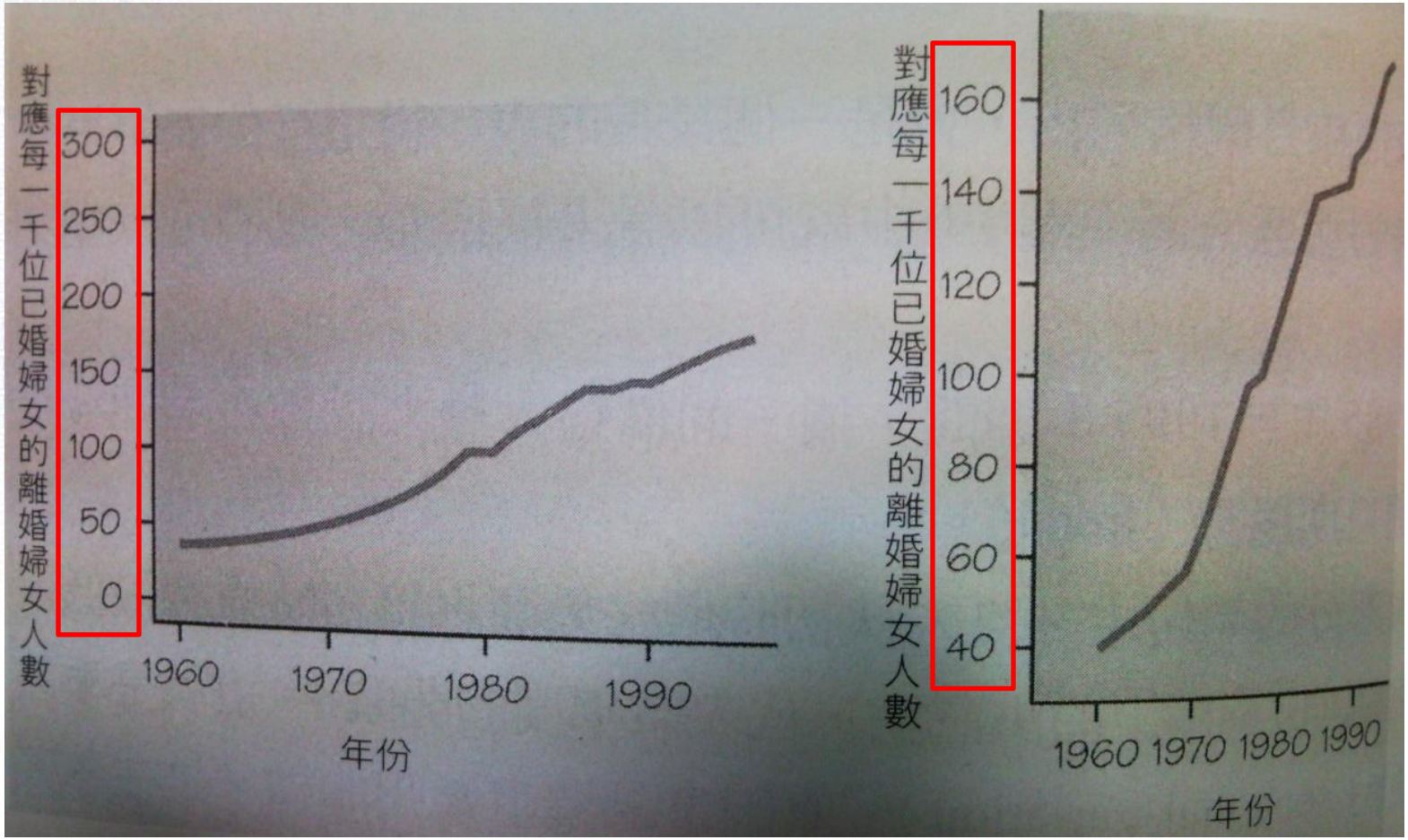
線圖

- 線圖（Line graphs）可以顯示出變數隨時間所產生的變化。時間刻度標示在橫軸上，而變數的刻度於縱軸上。



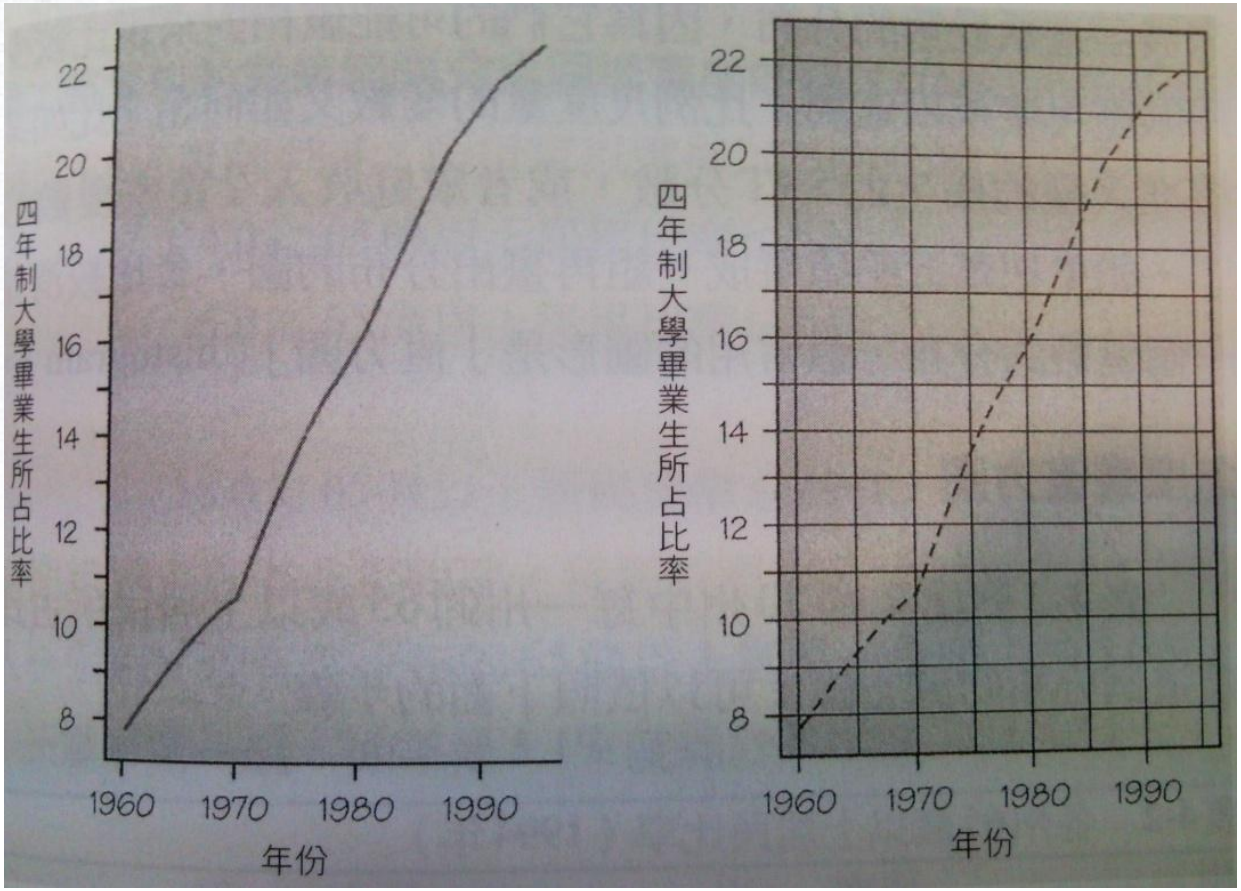
注意刻度大小！

- 圖形可以提供強烈的訊息，所以很容易誤導，於讀線圖時，需明確的判讀兩軸的刻度。



怎樣把圖畫好

- 清楚標示所畫的變數、單位、資料來源。
- 讓資料醒目，確實的抓住看圖者的注意力，而非背景。



如何畫直方圖

Step1.

- 將資料的數值範圍分成同樣寬度的組

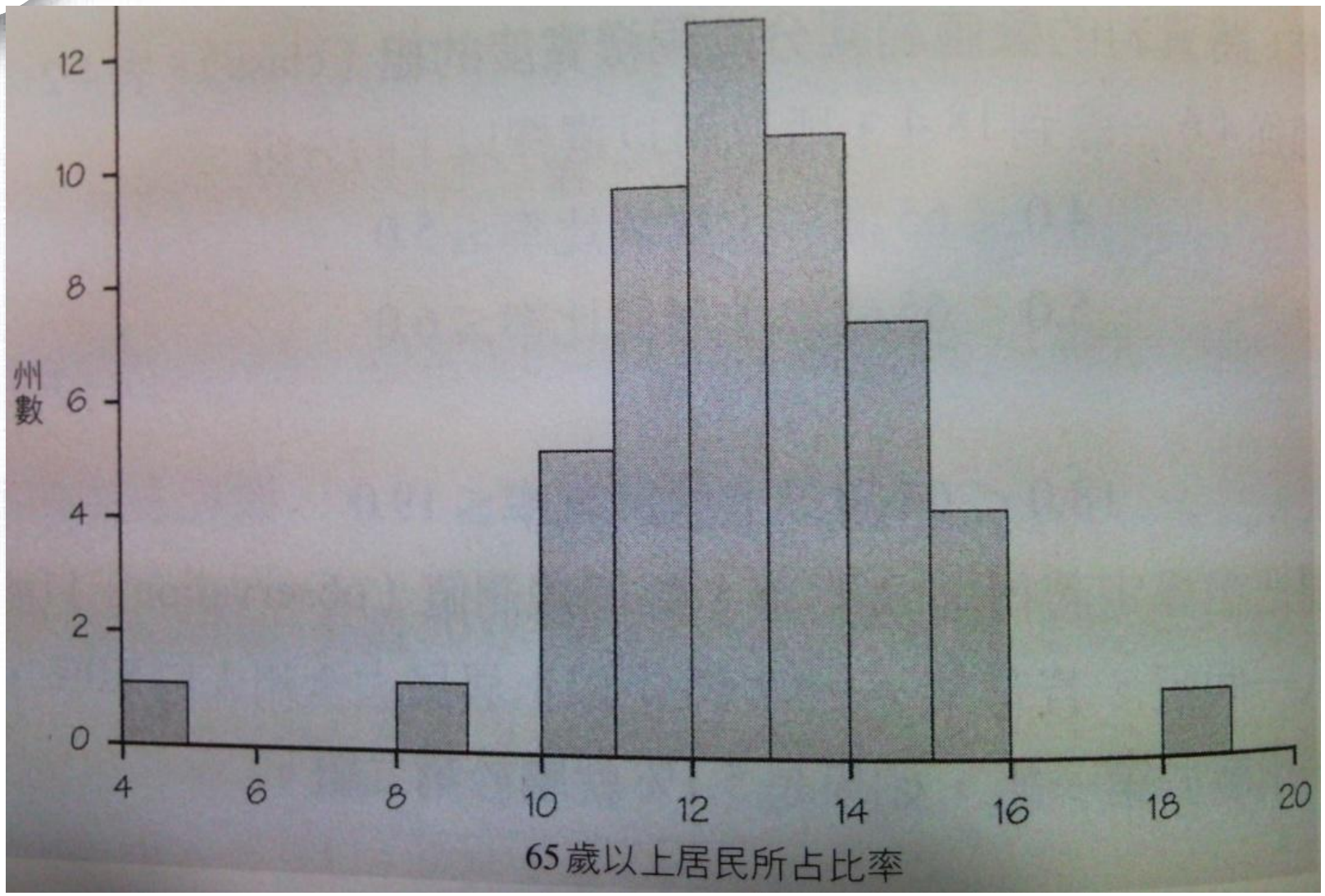
Step2.

- 數計每組中觀測值的個數

Step3.

- 畫直方圖

如何畫直方圖 (Cont.)



直方圖 vs. 長條圖

直方圖與長條圖之差異

- 直方圖的底部刻度都間隔同樣的單位數；長條圖無刻度。
- 直方圖裡的長條寬度有意義；長條圖則無意義。
- 直方圖中的長條互相鄰接，沒有空隙（除非有組計數為零）。

解釋直方圖

型態（pattern）及偏差（deviation）

- 在任何一組資料的圖形裡，我們要找的是一般型態，以及有異於一般型態的顯著偏差。

離群值

離群值（outlier）

- 一組資料的任何圖形之離群值，是指落在圖形一般型態之外的觀測值。

分布的一般型態

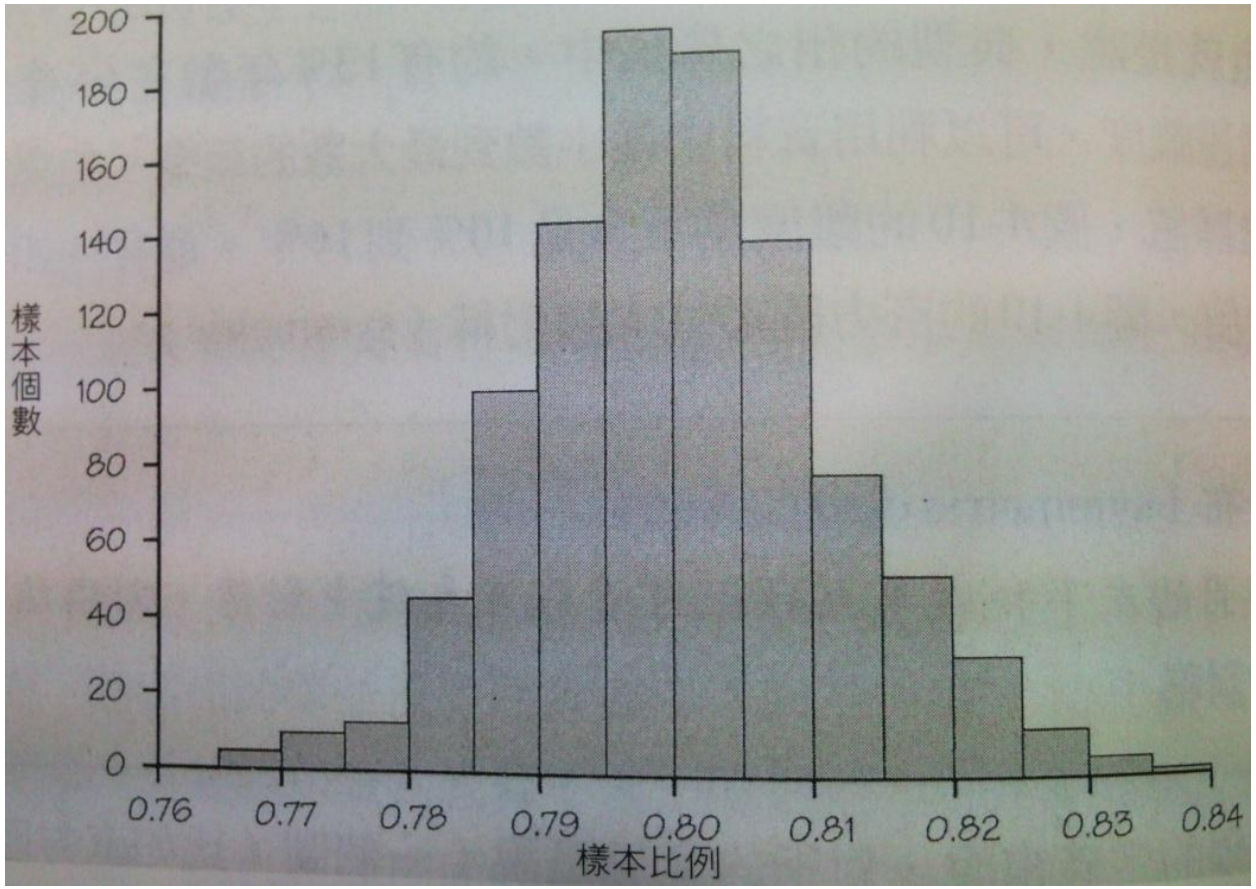
要描述分布的一般型態：

- 找出中心（center）及離度（spread）
- 看看該分布是否有可以用簡單的話描述之形狀。

對稱分佈

對稱分布（symmetric distribution）：

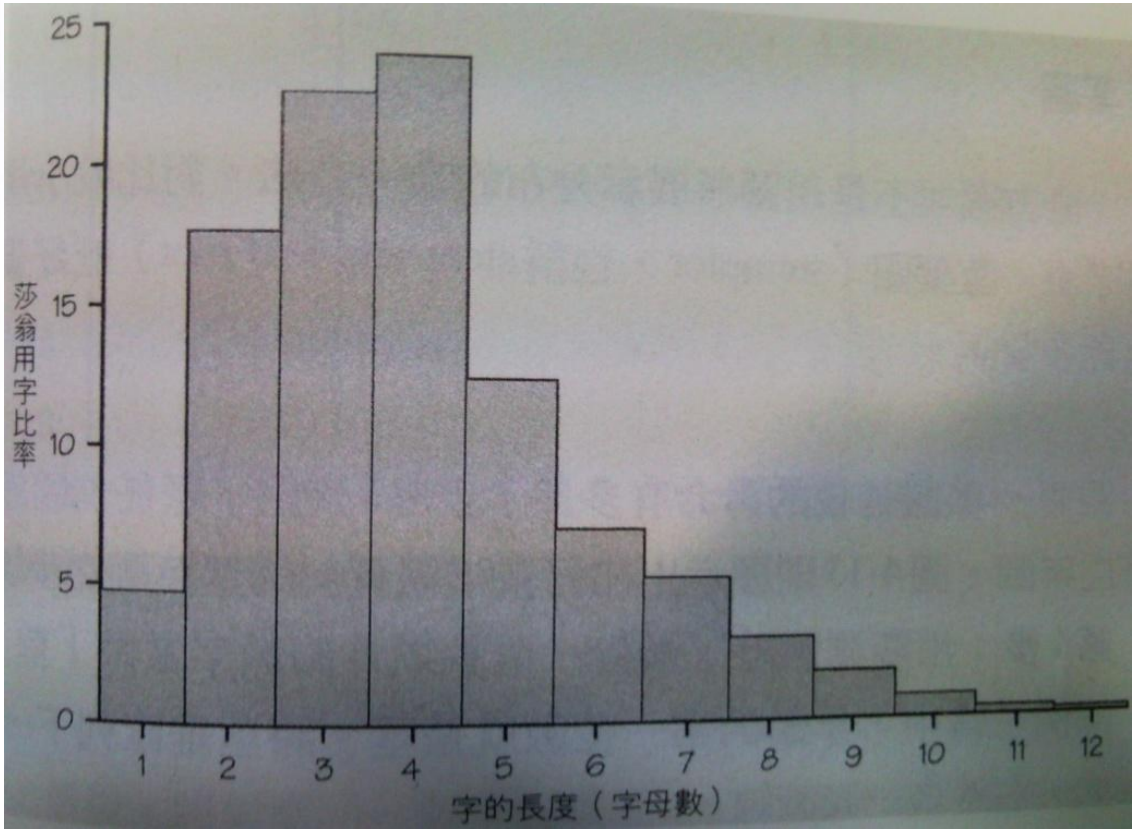
- 若直方圖的左半和右半大致上可看成互為鏡中影像，則稱該分佈為**對稱**。



偏斜分佈

偏斜分布（skewed distribution）：

- 若直方圖的右邊延伸比左邊多，則是右偏；若直方圖的左邊延伸比右邊多，則是左偏。



莖葉圖 (1/3)

莖葉圖（stemplot; stem-and-leaf plot）：

- 直方圖並不是用圖形展示分佈的唯一方法。對比較小規模的資料，莖葉圖既好畫也呈現出更多資訊。

莖葉圖之步驟：

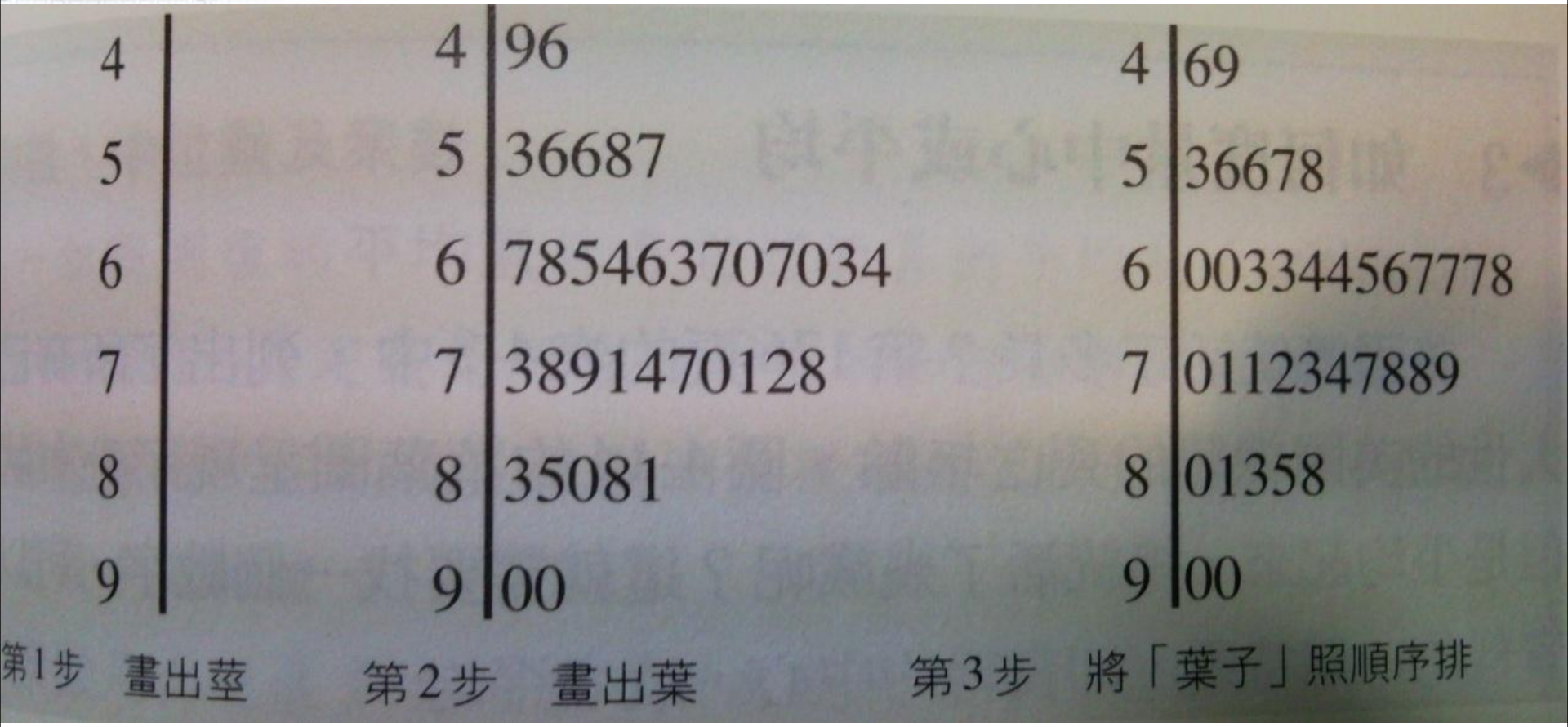
- Step1. 用觀測值除了最後一位數以外的數字當做”莖”。將莖垂直列下，由小到大，右邊畫一條直線。
- Step2. 將變數的個位數，寫在恰當的莖的右邊當做”葉子”。莖可以是任何位數，但葉只能是一位數。
- Step3. 最後，將對應同一個莖的葉子，由小到大重新排列。

莖葉圖 (2/3)

表4-3 美國歷任總統死亡年齡

華盛頓	67	費爾莫爾	74	老羅斯福	60
亞當斯	90	皮爾斯	64	塔夫特	72
傑佛遜	83	布坎南	77	威爾遜	67
麥迪遜	85	林肯	56	哈定	57
門羅	73	約翰遜	66	柯立芝	60
亞當斯	80	格蘭特	63	胡佛	90
傑克遜	78	海斯	70	小羅斯福	63
范布倫	79	加菲爾德	49	杜魯門	88
哈里森	68	阿瑟	56	艾森豪	78
太勒	71	克利夫蘭	71	甘迺迪	46
波克	53	哈里遜	67	詹森	64
泰勒	65	麥金萊	58	尼克森	81

莖葉圖 (3/3)


4
5
6
7
8
9
4 96
5 36687
6 785463707034
7 3891470128
8 35081
9 00
4 69
5 36678
6 003344567778
7 0112347889
8 01358
9 00
第1步 畫出莖
第2步 畫出葉
第3步 將「葉子」照順序排

如何度量中心或平均

- **平均數：**

  算數平均數，所有觀測值的合除以觀測值的個數。

- **中位數：**

  中間的值，把觀測值從小到大排列之後，中位數是最中間的值。

- **眾數：**

  出現最多次的值。

常用之度量中心

- 平均數：

  Ex. 美國大學測驗ACT（American College Testing）數學分數的平均。

- 中位數：

  Ex.一家之主為大學畢業生的家庭，收入中位數為年薪56,116美元；而一家之主在中學就輟學的家庭，收入中位數只有年薪28,700美元。

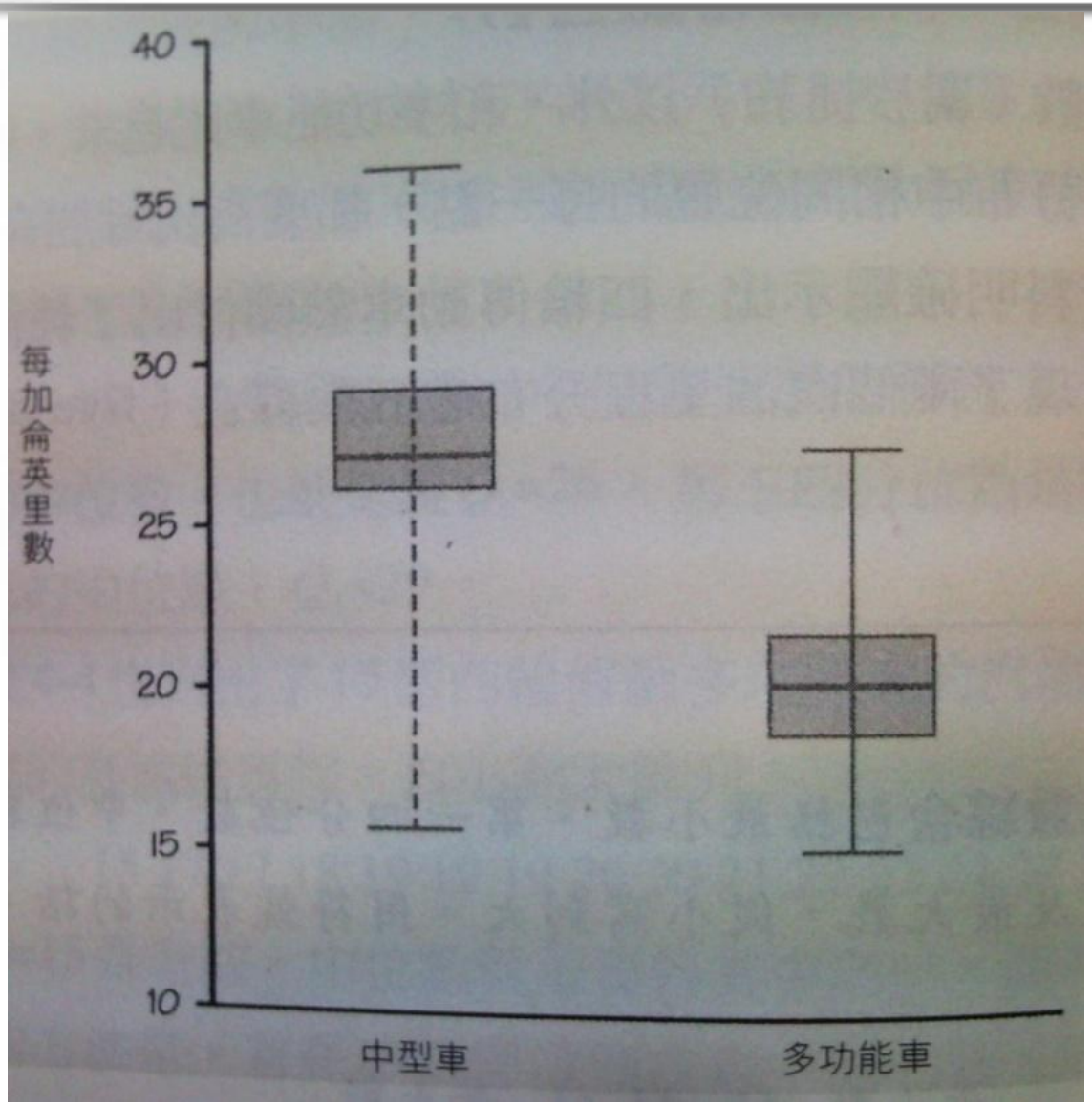
- 眾數：

  Ex.民國九十二年全台灣發生地震次數最多的縣市是花蓮縣。

如何度量離度-四分位數

- 將觀測值從小排到大，並在排好序的序列中找出中位數。
- 第一四分位數Q1，是中位數左邊所有數字的中位數。
- 第三四分位數Q3，是中位數右邊所有數字的中位數。
- 五數綜合：包括最小數、第一四分位數、中位數、第三四分位數及最大數

盒圖（Boxplot）



如何度量離度-標準差

- 一組觀測值的變異數s^2是所有觀測值距平均數的離差之平方的平均數，若用符號表示，n個觀測值$x_1, x_2, \ldots x_n$的變異數是

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

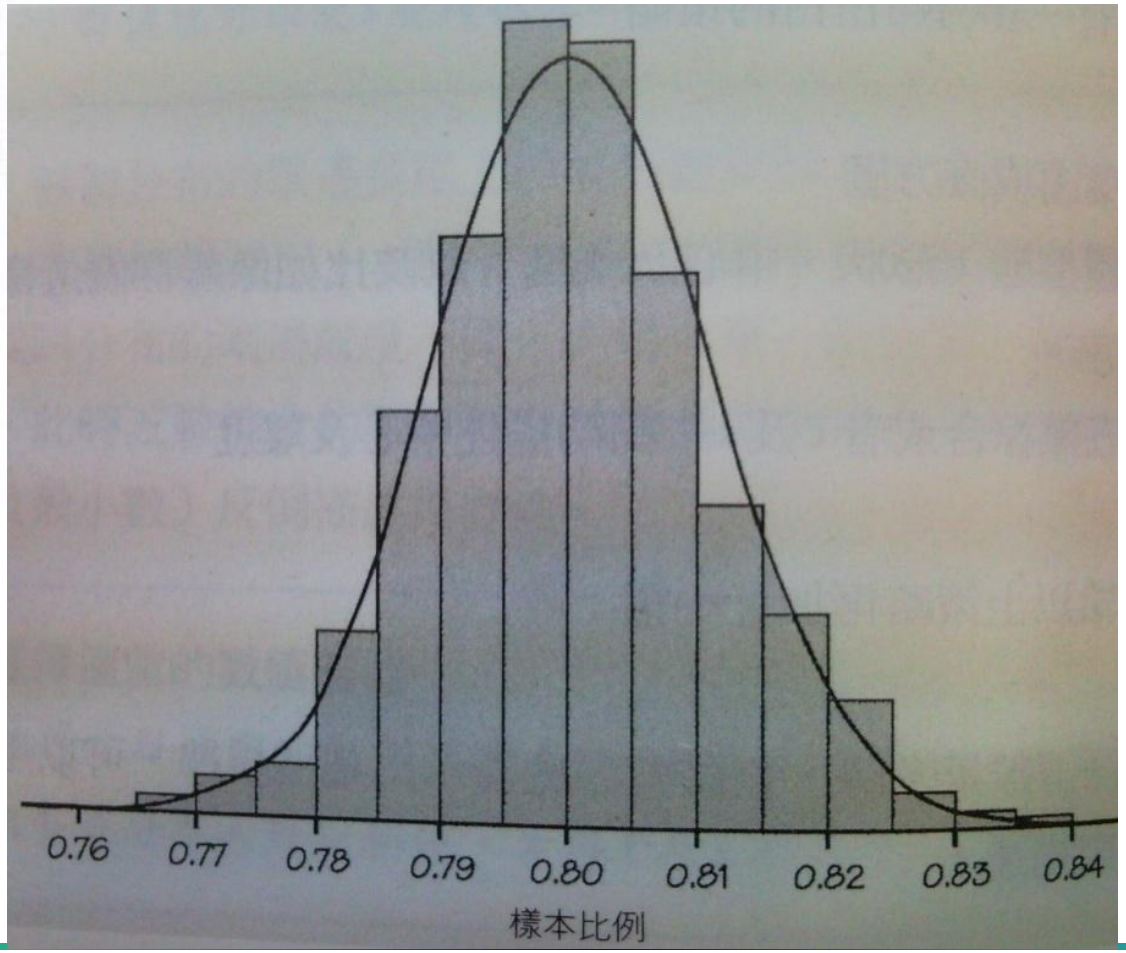
- 標準差s是變異數s^2的正平方根

標準差的特質

- 標準差度量以平均數為中心的離度，只有在用標準差來描述分布中心時，才可以用標準差來描述離度。
- 只有在沒有離度的時候，標準差才會等於零。而只有在所有觀測值都有相同的值時，這種情況才會發生。
- 標準差和平均數一樣，都會被極端值嚴重影響。少少幾個離群值就可以使標準差變的很大。
- 描述偏斜分布時，用五數綜合通常好過用平均數及標準差。只有分布大致對稱時，才用平均數及標準差。

常態分佈

- 曲線分布：有時觀測值數量多時，整體型態會顯示出某種規律，可以用平滑曲線來描述。



常態分佈 (Cont.)

- 沒有一組真實的資料是可以由一條密度曲線來完完全全描述的，密度曲線是理想的型態。
- 密度曲線是百分之百對稱的，但實際資料只是大致對稱。

密度曲線的中心和離度

- 眾數是密度曲線的尖峰點，就是曲線最高峰所在的值。
- 中位數是等面積點，也就是也就是曲線底下一半的面積在其左，一半面積在其右的值。
- 平均數是平衡點，就是假如用實心材料切割出密度曲線的圖形來，這個圖形的平衡點所在的值。

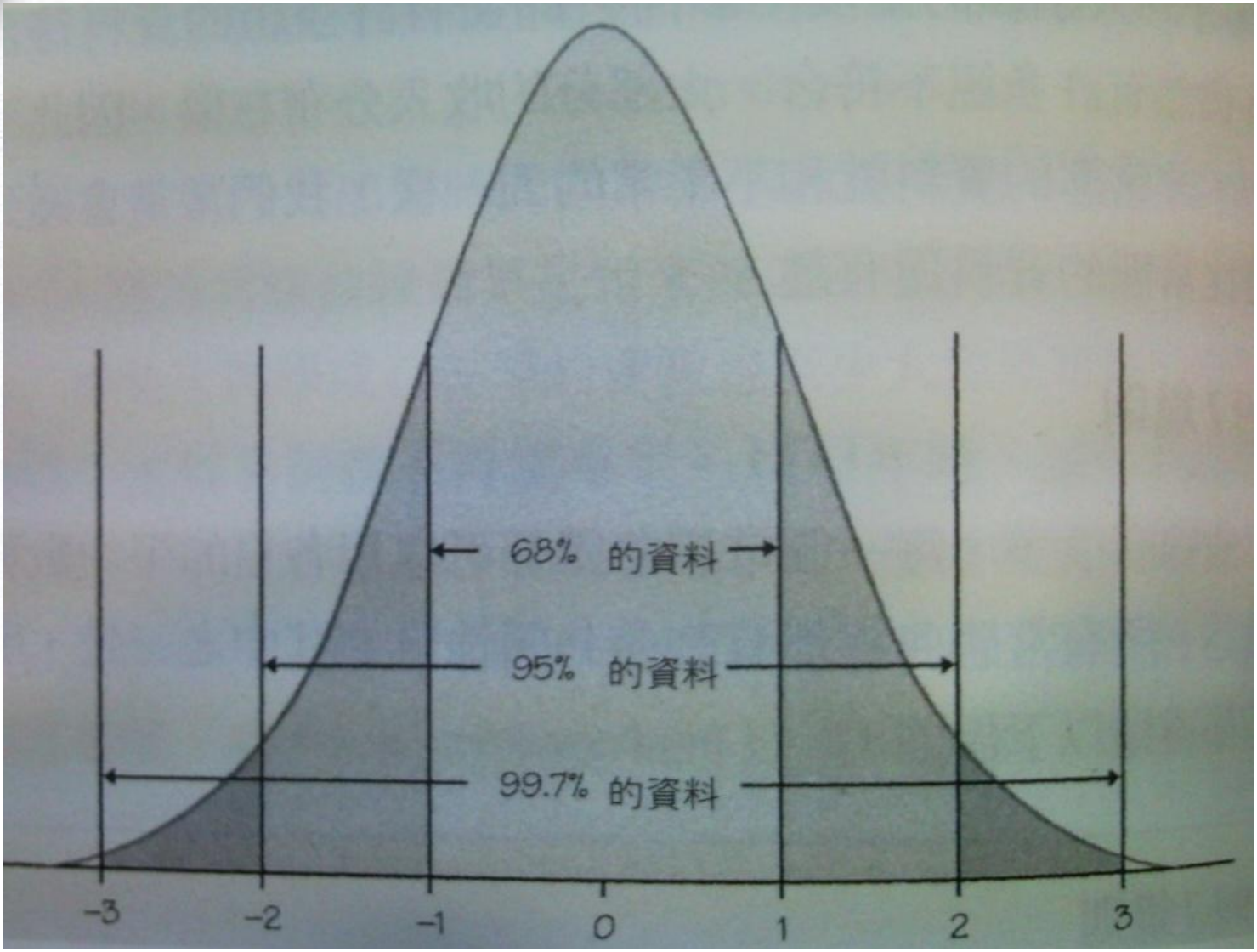
常態密度曲線的性質

- 只要給定平均數和標準差，就可以完全描述特定的常態曲線。
- 平均數決定分布的中心，位置在曲線的對稱中心。
- 標準差決定曲線的形狀，是從平均數到平均數左側或右側的曲線率轉換點的距離。

68-95-99.7規則

- 68%的觀測值落在距平均數一個標準差範圍內。
- 95%的觀測值落在距平均數二個標準差範圍內。
- 99.7%的觀測值落在距平均數三個標準差範圍內。

68-95-99.7規則 (Cont.)



高斯分布

- 任何變數，只要是許多相互獨立的小作用之和或平均，值的分布就會接近常態。
- 天文學家或測量員仔細重複度量同一數量時，會有小誤差，高斯用這些曲線來描述這些小誤差。
- 十九世紀的大部分時間中，常態曲線叫做誤差曲線。
- 1783年時這些曲線第一次由美國邏輯學家皮爾斯稱作常態曲線。

標準計分

- 以平均數為零點，將觀測值以標準差為單位表示出來，得到的數值就叫做標準計分。

標準計分=(觀測值-平均數)/標準差

標準計分 (Cont.)

Ex.

ACT分數對應SAT分數。愛蓮娜在SAT的數學部分考了680分。SAT的分數遵循平均數為500、標準差為100的常態分布。傑若考了美國大學測驗ACT的數學部分得了27分。ACT分數是常態分布，平均數為18，標準差6。假設兩種測驗的評量標的差不多，誰的分數比較高？

Sol:

愛蓮娜的標準計分是： (680-500)/100=1.8

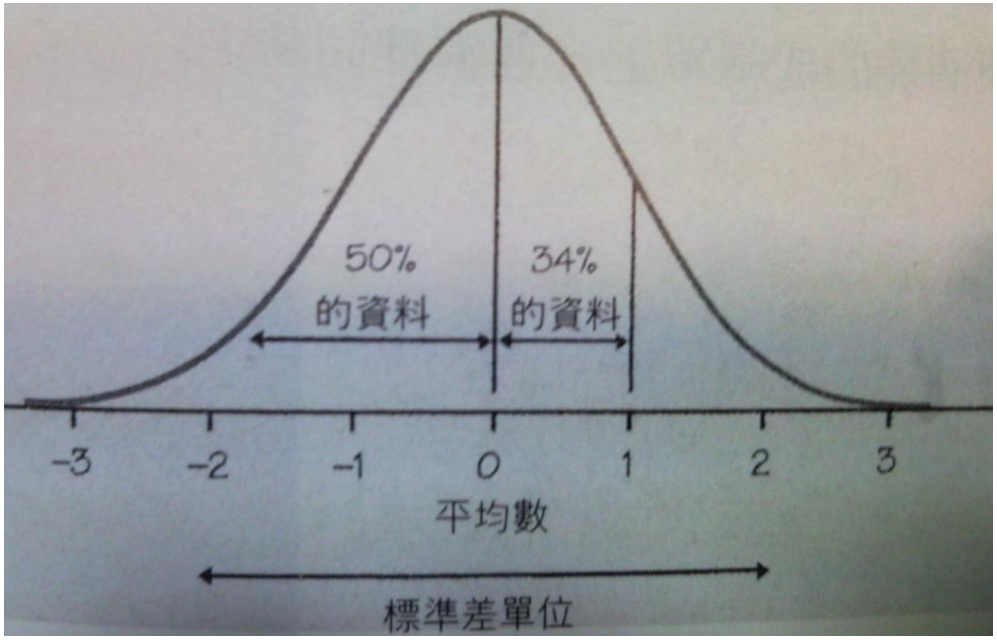
傑若的標準計分是： (27-18)/6=1.5

→愛蓮娜的分數在平均數之上1.8個標準差，而傑若只在平均數以上1.5個標準差，愛蓮娜考的比較好。

常態分布的百分位數

- 對常態分布來說，標準差提供了相當準確的比較，因為標準差可以直接轉換成百分位數。

- 一個分布的第c百分位數是一個值：小於第c個百分位數的觀測值，再全部觀測值所佔百分比為c，其餘的觀測值則比第c百分位數大。



Thanks for your attention