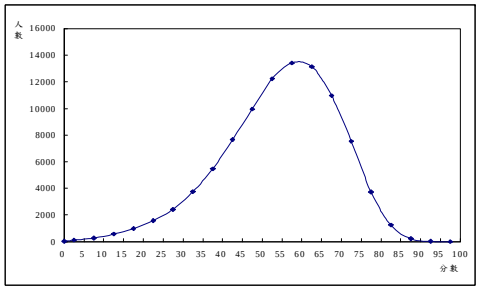
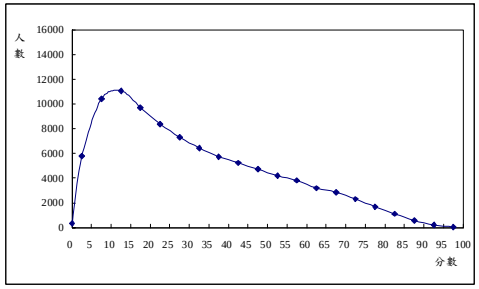
第 5 章　分析資料

以統計測量數來呈現

- 統計表、統計圖雖可用來顯示及描述統計資料的一般型態或分散的情形，但所提供的資訊只是簡潔的摘要性描述，不能提供精確的描述。
  - 精確具體的資料描述：用數字來描述資料的特性。
- 尤其是當我們在比較兩組資料的性質時，僅依賴統計圖表所做出來的比較會過於粗略，最好可以有精確的數字作為比較的基礎。
  - 例子：96 學年度大學指考國文科與英文科成績的比較。由統計圖可以看出，國文科成績分布較對稱(雖稍有左偏)，接近常態分配，英文科成績則有嚴重的右偏情況，顯示分數集中在較低分。整體看來，英文科的平均成績較國文科來的低。但這樣的比較過於粗略。

國文科成績的次數多邊圖　　英文科成績的次數多邊圖





國文科與英文科成績的統計測量數

	國文科	英文科
平均數	54.4422	31.0917
變異數	211.4321	464.9560
標準差	14.5407	21.5628
眾數	57.5000	12.5000
中位數	53.4756	23.7921

如果有統計測量數，我們就可做一些較精確的描述

(1) 平均數：國文科平均成績 54.44，高於英文科的平均 31.09。

(2) 變異數(標準差)：英文成績的變異程度較大(變異數為 464.96，標準差為 21.56)。

(3) 眾數：英文考 12.5 分的最多，國文考 57.5 分的最多。

(4) 中位數：國文科有一半的學生分數低於 53.48 分，英文科有一半的學生分數低於 23.89 分。

- 本章將資料區分成
  - 分組資料：已經被分類或分組的資料。
  - 未分組資料：未經分類或分組的資料。

未分組資料中心位置的衡量

- 統計圖可觀察資料的中心位置(資料的共同趨勢)，但這只是粗略的描述，使用數字來描述會比較具體且精確。

平均數(mean)

- 平均數是衡量資料中心位置最重要的**測量數**(measure)。
- 平均數除了代表一組資料的平均水準外，亦可用來比較兩組或兩組以上資料的平均水準。
- 依平均方式的不同，可將平均數區分為
  - 算術平均數(arithmetic mean)
  - 幾何平均數(geometric mean)
  - 調和平均數(harmonic mean) [少用，不介紹]
- **算術平均數**：所有觀察值的總和除以觀察值的個數。
  - 母體平均數：假設 X 是我們所關心的變數，$x_1,...,x_N$ 等 N 個觀察值是變數 X 的母體資料，則母體平均數(μ)定義為

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- 樣本平均數：若 $x_1,...,x_n$ 等 n 個觀察值是變數 X 的樣本資料，則樣本平均數($\bar{X}$)定義為

$$\bar{X}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{x_1+x_2+\cdots+x_n}{n}$$

- 例子：銀行業與證券業的薪資哪個高？抽樣調查兩個行業各 7 個初任員工的薪資資料如下

	A	B	C	D	E	F	G	H
1	證券業	20	23	23	25	26	29	64
2	銀行業	26	27	27	28	30	32	33

這些是樣本資料。證券業 7 個初任員工的平均薪資為

$$\bar{X}=(20+23+23+25+26+29+64)/7=30$$

銀行業 7 個初任員工的平均薪資為

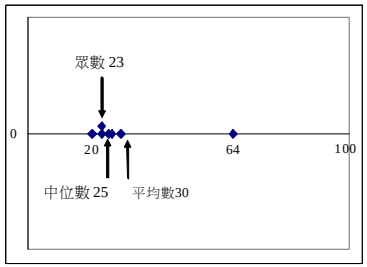
$$\bar{X}=(26+27+27+28+30+32+33)/7=29$$

從平均薪資看來，證券業的薪資比銀行業高。

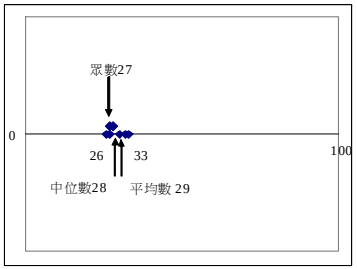
- 補充：加總(summation)符號
  - 若 $x_1,...,x_n$ 代表變數X的n個觀察值[n個不同數字]，則 $\sum_{i=1}^{n}x_i$ 表示這n個觀察值的總和。以證券業7個初任員工的薪資為例，$n=7$，令 $x_1=20$、$x_2=23$、$x_3=23$、$x_4=25$、$x_5=26$、$x_6=29$、$x_7=64$，則 $\sum_{i=1}^{7}x_i$ 代表這7個數字的總和。
  - 若 $x_1,...,x_n$ 代表變數X的n個觀察值，$y_1,...,y_n$ 代表變數Y的n個觀察值，且a與b是固定常數(不會因觀察值下標i變動而改變)，我們可得到底下幾個加總的性質
    - $\sum_{i=1}^{n}a=\underbrace{a+\cdots+a}_{n個}=na$
    - $\sum_{i=1}^{n}ay_i=ay_1+ay_2+\cdots+ay_n$
      $=a(y_1+y_2+\cdots+y_n)=a\sum_{i=1}^{n}y_i$

- $\sum_{i=1}^{n}(ay_i+bx_i)=(ay_1+bx_1)+(ay_2+bx_2)+\cdots+(ay_n+bx_n)$
  $=a(y_1+y_2+\cdots+y_n)+b(x_1+x_2+\cdots+x_n)$
  $=a\sum_{i=1}^{n}y_i+b\sum_{i=1}^{n}x_i$

- 算數平均數的特質
  - 資料的平衡點：平均數左邊的觀察值與平均數之距離的總和，等於平均數右邊的觀察值與平均數之距離的總和。



證券業的平均月薪之點圖



銀行業的平均月薪之點圖

  - 各觀察值與平均數間的差的總和等於零，亦即

$$\sum_{i=1}^{N}(x_i-\mu)=0 \quad 或 \quad \sum_{i=1}^{n}(x_i-\bar{X})=0$$

[證明] 因 $\mu=\frac{1}{N}\sum_{i=1}^{N}x_i$，所以 $N\mu=\sum_{i=1}^{N}x_i$，故

$$\sum_{i=1}^{N}(x_i-\mu)=\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\mu=\sum_{i=1}^{N}x_i-N\mu=\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}x_i=0$$

又因 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$，所以 $n\bar{X}=\sum_{i=1}^{n}x_i$，故

$$\sum_{i=1}^{n}(x_i-\bar{X})=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{X}=\sum_{i=1}^{n}x_i-n\bar{X}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0$$

  - 各觀察值與平均數之差的平方和最小：若 A 代表任意實數，則使得 $\sum_{i=1}^{N}(x_i-A)^2$ 最小的 A 為 μ，使得 $\sum_{i=1}^{n}(x_i-A)^2$ 最小的 A 為 $\bar{X}$。亦即

$$\sum_{i=1}^{N}(x_i-\mu)^2\le\sum_{i=1}^{N}(x_i-A)^2$$

$$\sum_{i=1}^{n}(x_i-\bar{X})^2\le\sum_{i=1}^{n}(x_i-A)^2$$

[證明]

$$\sum_{i=1}^{N}(x_i-A)^2=\sum_{i=1}^{N}(x_i-\mu+\mu-A)^2$$
$$=\sum_{i=1}^{N}[(x_i-\mu)^2+(\mu-A)^2+2(\mu-A)(x_i-\mu)]$$
$$=\sum_{i=1}^{N}(x_i-\mu)^2+\sum_{i=1}^{N}(\mu-A)^2+2(\mu-A)\sum_{i=1}^{N}(x_i-\mu)$$
$$=\sum_{i=1}^{N}(x_i-\mu)^2+\sum_{i=1}^{N}(\mu-A)^2+0$$
$$\geq\sum_{i=1}^{N}(x_i-\mu)^2$$
$$\sum_{i=1}^{n}(x_i-A)^2=\sum_{i=1}^{n}(x_i-\bar{X}+\bar{X}-A)^2$$
$$=\sum_{i=1}^{n}[(x_i-\bar{X})^2+(\bar{X}-A)^2+2(\bar{X}-A)(x_i-\bar{X})]$$
$$=\sum_{i=1}^{N}(x_i-\bar{X})^2+\sum_{i=1}^{N}(\bar{X}-A)^2+2(\bar{X}-A)\sum_{i=1}^{N}(x_i-\bar{X})$$
$$=\sum_{i=1}^{N}(x_i-\bar{X})^2+\sum_{i=1}^{N}(\bar{X}-A)^2+0$$
$$\geq\sum_{i=1}^{N}(x_i-\bar{X})^2$$

- 平均數的優點是考慮到每一個觀察值[因平均數是所有觀察值的總和除以觀察值個數，故用到所有觀察值，且每一觀察值的重要性相同]，缺點是易受極端值(離群值)的影響[如證券業薪資的例子中，有一個 64(千元)的極端值，若去除掉該極端值，則證券業平均月薪僅有 24.3(千元)，可見極端值對平均數有極大影響]。
- 平均數可進行代數(加減乘除)演算[習題 5.13]。
- 可對觀察值予以加權[加權平均數]

● **加權算術平均數(weighted arithmetic mean)**

- 在現實世界中，觀察值的重要性不見得相同。例如：學生成績、物價指數、股價指數的計算中，因為學科的成績與分數、商品的產量與價格、股票之發行量與股價均不相同，故不能等同看待每一學科、商品、股票，每個觀察值對個人或整個經濟的影響亦不相同。

- 在計算學生成績、物價指數、股價指數時，採取加權的方式，以學分、商品數量、股票市值為**權重**來計算平均成績或指數。亦即當觀察值重要性不一樣時，可給予每個觀察值一個權數，用來代表其重要性，然後再計算其平均數。
- 以W_i代表各觀察值的加權數，則加權算術平均數之計算為

母體： $\mu_w=\dfrac{\sum_{i=1}^{N}W_ix_i}{\sum_{i=1}^{N}W_i}$　　樣本： $\bar{X}_w=\dfrac{\sum_{i=1}^{n}W_ix_i}{\sum_{i=1}^{n}W_i}$

我們通常令$w_i=W_i\Big/\sum_{i=1}^{N}W_i$ (或$w_i=W_i\Big/\sum_{i=1}^{n}W_i$)，而將加權算術平均數改寫為

母體： $\mu_w=\sum_{i=1}^{N}w_ix_i$　　樣本： $\bar{X}_w=\sum_{i=1}^{n}w_ix_i$

此時的w_i可解釋為每一觀察值之權數比例(重要性比例)

$\sum_{i=1}^{N}w_i=\sum_{i=1}^{N}W_i\Big/\sum_{i=1}^{N}W_i=1$　且　$\sum_{i=1}^{n}w_i=\sum_{i=1}^{n}W_i\Big/\sum_{i=1}^{n}W_i=1$

- 例子：學期總成績。某個學生的學期成績單如下

	A	B	C	D	E
1	科目	學分數	學期成績	加權成績	
2	國文	3	86	258	
3	英文	3	87	261	
4	歷史	2	95	190	
5	體育	1	87	87	
6	微積分	3	82	246	
7	經濟學原理	4	86	344	
8	會計學	3	89	267	
9	社會心理學	3	90	270	
10	工程發展與社會變遷	2	91	182	
11		24		2105	87.71

其學期平均成績(加權算術平均數)可計算如下

$\mu_w=\sum_{i=1}^{N}W_ix_i\Big/\sum_{i=1}^{N}W_i=(3\times86+3\times87+\cdots+2\times91)/24=87.71$

或是以$\frac{3}{24}$、$\frac{3}{24}$、$\frac{2}{24}$、$\frac{1}{24}$、$\frac{3}{24}$、$\frac{4}{24}$、$\frac{3}{24}$、$\frac{3}{24}$等學分比例為權重，計算為

$\mu_w=\sum_{i=1}^{N}w_ix_i=\frac{3}{24}\times86+\frac{3}{24}\times87+\cdots+\frac{2}{24}\times91=87.71$

- **樣本的幾何平均數**：若資料為等比數列，如國民生產毛額成長率、物價上漲率、投資報酬率等，應以幾何平均數來代表該等資料的中心位置。
  - 若 $x_1, x_2, ..., x_n$ 為變數 X 的 n 個樣本觀察值，且均為正數，則樣本幾何平均數為

$$\bar{g} = \sqrt[n]{x_1 \cdot x_n \cdots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

- **幾何平均數的性質：**

$$\sqrt[n]{\prod_{i=1}^{n}(x_i / y_i)} = \sqrt[n]{\prod_{i=1}^{n} x_i} \Big/ \sqrt[n]{\prod_{i=1}^{n} y_i}$$

  - 例如：X 為國民所得，Y 為人口數，則 X/Y 為平均每人所得，要求算其平均成長率可以計算 X/Y 的幾何平均數，或分別計算 X 與 Y 的幾何平均數，再將兩個幾何平均數相除。
  - 幾何平均數適合衡量等比數列的中央位置，但不易進行統計推論。

  - 例子：下表的資料是台積電 90~96 年年底的股票收盤價

年度	台積電股價	變動比(P_t / P_{t-1})
90	77.74	
91	67.42	0.867
92	56.42	0.837
93	52.36	0.928
94	54.08	1.033
95	61.34	1.134
96	65.52	1.068

則台積電股價變動率的幾何平均數為

$$\bar{g}_{台積電} = (0.867 \times 0.873 \times 0.928 \times 1.033 \times 1.134 \times 1.068)^{\frac{1}{6}} = 0.9719$$

幾何平均投資報酬率為 $R_{台積電} = 0.9719 - 1 = -0.0281 = -2.81\%$

台積電股價變動率的算術平均數為

$$\bar{X}_{台積電} = (0.867 + 0.873 + 0.928 + 1.033 + 1.134 + 1.068)/6 = 0.9778$$

算術平均投資報酬率為 $0.9778 - 1 = -0.0221 = -2.21\%$，低估了真正的平均投資報酬率 -2.81%。

- **幾何平均數的投資報酬率**
  - 由於我們在討論投資報酬率時，會考慮複利(再投資)的效果，因此用幾何平均數來計算平均投資報酬率較恰當。
  - 假設有一項投資持續了 n 期，其每一期的投資報酬率分別為 $R_1, R_2, ..., R_n$，則幾何平均數的投資報酬率為

$$\bar{G} = [(1+R_1)(1+R_2)\cdots(1+R_n)]^{\frac{1}{n}} - 1$$

  - 例子：假設某人將 1000 萬投資在股票上，在接下來的兩年年年底，其財富變化如下表所示，表中我們亦計算出每年的離散報酬率[(期末金額-期初金額)/期初金額]

時間點	財富總額	每年報酬率
第 0 年年底	1000 萬	
第 1 年年底	500 萬	-50% [=(500-1000)/1000]
第 2 年年底	750 萬	50% [=(750-500)/500]

若運用算術平均數來計算每年的平均投資報酬率，會得到

$$\bar{X} = (-50\% + 50\%)/2 = 0\% \quad \text{[沒有虧損、沒有獲利]}$$

但是很明顯地，這筆投資是虧損的(兩年累積虧損 25%)。若考慮複利的狀況，計算每年平均投資報酬率($\bar{g}$)的正確方法應為

$$(1+\bar{g})(1+\bar{g}) = (1+R_1)(1+R_2)$$

$$\Rightarrow \bar{g} = [(1+R_1)(1+R_2)]^{\frac{1}{2}} - 1$$

$$\Rightarrow \bar{g} = [(1-50\%)(1+50\%)]^{\frac{1}{2}} - 1 = -0.134 = -13.4\%$$

亦即，平均而言，每年的投資報酬率是 -13.4%，兩年累計虧損了 25%。

一般化

$$\underbrace{(1+\bar{g})(1+\bar{g})\cdots(1+\bar{g})}_{n\text{個}} = (1+R_1)(1+R_2)\cdots(1+R_n)$$

$$\Rightarrow \bar{g} = [(1+R_1)(1+R_2)\cdots(1+R_n)]^{\frac{1}{n}} - 1$$

中位數(median)

- **中位數**：將觀察值依數值大小順序排列後，居於中央的那一個數值稱為中位數。[僅適用於具有順序的資料，不適用於類別資料]
  - 所有的觀察值至少有一半(1/2)大於等於中位數，而且至少有一半(1/2)小於等於中位數。換句話說，『大於等於中位數的觀察值』與『小於等於中位數的觀察值』都至少有一半。
  - 當資料中有極端值存在時，算術平均數不是一個良好的指標，此時可用中位數來代表資料的中心位置。
  - 符號：以 m_e 來代表中位數。
- 中位數的求算：若變數 X 的 N 個觀察值，可將其由小到大排序為 $x_1 < x_2 < \cdots < x_N$ [不管是母體資料或樣本資料，計算方法相同]
  - 若 N 為奇數，則中位數 m_e 為第 $\frac{N+1}{2}$ 個觀察值的數值。
  - 若 N 為偶數，則中位數 m_e 為第 $\frac{N}{2}$ 個與第 $\frac{N}{2}+1$ 個觀察值的平均數。

- 例子：銀行業與證券業的薪資哪個高？(以中位數比較)
  - 稍早我們提過，證券業的薪資中有一個極端值 6.4 萬，因此證券業的平均薪資會受此極端值影響，因而失去了代表性。
  - 中位數：由於樣本個數 $n=7$，所以中位數應位於第 $\frac{n+1}{2}=\frac{7+1}{2}=4$ 個觀察值；將銀行業與證券業的薪資資料由小至大排序

    證券業　20　23　23　25　26　29　64

    銀行業　26　27　27　28　30　32　33

    所以證券業薪資的中位數為 25(千元)，銀行業薪資的中位數為 28(千元)。若根據中位數來比較，銀行業的薪資比證券業高。
  - 由於證券業的薪資資料中存在極端值，故其平均數(3 萬)與中位數(2.5 萬)差異較大，去除該極端值後所計算的平均值(2.43 萬)與中位數較接近。銀行業薪資資料並無極端值存在，故其平均數(3 萬)與中位數(2.5 萬)較接近。

眾數(mode)

- 眾數：眾數是指觀察值中出現次數最多的那一個數值或類別。
- **眾數的性質**
  - 不受極端值的影響。
  - 對觀察值的個數或數值變化的反應不靈敏。
  - 眾數可能有很多個，也可能一個也沒有，因此眾數比中位數及平均數較少使用。
- 例子：
  - 在銀行業與證券業的薪資的例子中，證券業薪資的眾數為 2.3 萬，銀行業薪資的眾數為 2.7 萬。
  - 類別資料的眾數其實就是佔全部比率最多的類別。
  - 若為計量尺度資料，通常以分組資料次數最多的那一組的組中點作為眾數。例如：指考國文科的眾數為 57.5，英文科的眾數為 12.5。

中心位置各統計測量數的比較與選擇

統計測量數	優　點	缺　點
算術平均數	1.資料的重心。資料無極端值或偏態時，具代表性。 2.適合代數演算 3.考慮所有觀察值，敏感度高。 4.觀察值與平均數差平方和最小 5.適合統計推論的工作	1.若有極端值存在時則不具代表性 2.資料如為偏態，則代表性較差。
幾何平均數	1.適合等比資料 2.敏感度高	1.不適合一般資料 2.不適合作統計推論
中位數	1.適用於有極端值的資料 2.適用於偏態資料 3.觀察值與中位數絕對差和最小 4.可做無母數統計推論	1.不適合代數演算 2.對觀察值敏感性低 3.不易進行母數統計推論
眾數	1.適用於有極端值的資料 2.適用於偏態資料 3.適用於質的資料	1.可能不止一個或不存在 2.敏感性低 3.不能做統計推論

未分組資料等分位置的衡量

四分位數(quartiles)

- 四分位數是將順序資料分成四等分數值的分位數。四分位數有第1、第2、第3三個四分位數。
  - 四分位數的基本想法：將觀察值從最小值到最大值分成四等分，每一等分各佔25%的分位數(分割點)。
  - 若以Q_i $(i=1,2,3)$來表示第i個四分位數，則至少有$\frac{i}{4}$的觀察值小於等於Q_i，而且至少有$1-\frac{i}{4}$的觀察值大於等於Q_i。
- 四分位數的求算：假設觀察值個數有n個，且已經由小至大排序
  - 第1四分位數：計算$K=\frac{n}{4}$，若K為整數，則Q_1為第K個與第$K+1$個觀察值的平均數；若K不是整數，則將K無條件進位(設其為整數M)，則Q_1為第M個觀察值的數值。
  - 第2四分位數：Q_2就是中位數。

  - 第3四分位數：計算$K=\frac{3n}{4}$，若K為整數，則Q_3為第K個與第$K+1$個觀察值的平均數；若K不是整數，則將K無條件進位(設其為整數M)，則Q_1為第M個觀察值的數值。
- 例子：假設有9個高中男生的體重(單位：公斤)如下[已經排序過]
  55　57　62　63　66　67　73　75　81
  - 四分位數分別為

    Q_1：$K=\frac{n}{4}=\frac{9}{4}=2.25$，無條件進位為3，故第3個觀察值為$Q_1$，亦即$Q_1=62$。[有$\frac{3}{9}$的觀察值(已超過25%)小於等於$Q_1=62$，有$\frac{7}{9}$的觀察值(已超過75%)大於等於$Q_1=62$]

    Q_2：因觀察值個數為奇數，故中位數為第$\frac{n+1}{2}=\frac{9+1}{2}=5$個觀察值66，因此$Q_2=66$。

    Q_3：$K=\frac{3n}{4}=\frac{3\cdot 9}{4}=6.75$，無條件進位為7，故第7個觀察值為$Q_3$，亦即$Q_3=73$。

十分位數(deciles)

- 十分位數是將順序資料的觀察值均分為十等份數值的分割數。
  - 十分位數有9個，將第i個十分位數記為D_i $(i=1,2,...,9)$，則至少有$\frac{i}{10}$的觀察值小於等於D_i，而且至少有$1-\frac{i}{10}$的觀察值大於等於D_i。
- 十分位數的求算：假設觀察值個數有n個，且已經由小至大排序
  - 計算$K=\frac{n\cdot i}{10}$ $(i=1,2,...,9)$
  - 若K為整數，則取第K個與第$K+1$個觀察值的平均數為第i個十分位數D_i。
  - 若K不是整數，則將K無條件進位(設其為整數M)，取第M個觀察值為第i個十分位數D_i。
- 例子：某班級18個學生的學期成績如下(已依大小排序過)

78	79	80	81	82	83	83	84	84
85	86	87	88	89	90	91	92	95

  - D_5：$K=\frac{n\cdot i}{10}=\frac{18\cdot 5}{10}=9$，故取第9個與第10個觀察值的平均數為第5個十分位數，即$D_5=\frac{84+85}{2}=84.5$
  - D_8：$K=\frac{n\cdot i}{10}=\frac{18\cdot 8}{10}=14.4$，無條件進位為15，故第15個觀察值為第8個十分位數，即$D_8=90$。
  - 其餘分位數分別為：$D_1=79$、$D_2=81$、$D_3=83$、$D_4=84$、$D_6=86$、$D_7=88$、$D_9=91$，請自行練習。

百分位數(percentiles)

- 百分位數是將順序資料的觀察值均分為一百等分數值的分割數。
  - 百分位數有99個，第i個百分位數記為P_i $(i=1,2,...,99)$，至少有$\frac{i}{100}$的觀察值小於等於P_i，而且至少有$1-\frac{i}{100}$的觀察值大於等於P_i。
- 百分位數可讓我們立即得知，比某個觀察值大或小的資料之約略比率。

- 例如：若某個觀察值介於第 85 個百分位數與第 86 個百分位數間，即可知『比該觀察值小的資料佔了 85%，比該觀察值大的資料佔了 15%』。
- 當資料個數太少時，求算百分位數是一件不切實際的事，因為會大部分的百分位數都是重複的。此時若依然想知道某觀察值在資料當中的相對位置，可利用百分位比公式。
- **百分位比公式**：欲計算某個特定的觀察值在一群資料中的相對位置，可依底下公式計算小於該觀察值的資料比率

$$\text{百分位比}=\frac{\text{小於該觀察值的資料個數}+0.5}{\text{觀察值的總個數}}\times 100\%$$

- 例子：12 個大學男生鞋子的尺寸(單位：英吋)為

13　11　10　13　11　10　8　12　9　9　8　9

若我們想知道 12 英吋的百分位比為何，可先將資料排序為

8　8　9　9　9　10　10　11　11　12　13　13

比 12 英吋小的觀察值個數有 9 個，因此其百分位比為

$\frac{9+0.5}{12}\times 100\%=79.17\%$

根據 12 英吋的百分位比 79%，我們可以說『有 79%的大學男生鞋子尺寸小於 12 英吋，有 21%的大學男生鞋子尺寸小於 12 英吋』。

- 百分位數的求算：假設觀察值個數有 n 個，且已經由小至大排序
  - 計算 $K=\frac{n\cdot i}{100}$ $(i=1,2,...,99)$
  - 若 K 為整數，則取第 K 個與第 $K+1$ 個觀察值的平均數為第 i 個百分位數 P_i。
  - 若 K 不是整數，則將 K 無條件進位(設其為整數 M)，取第 M 個觀察值為第 i 個百分位數 P_i。

中位數、三種分位數間之關係

- $m_e=Q_2=D_5=P_{50}$ 、 $Q_1=P_{25}$ 、 $Q_3=P_{75}$ 、 $D_i=P_{10i}$

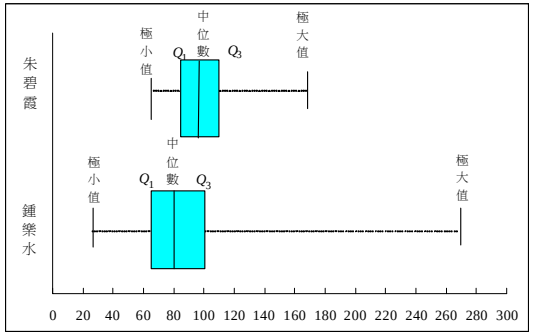
五數綜合－盒鬚圖分析法

- 從平均數、中位數、分位數，可知道資料的中心位置及資料等分位置的情形，但無法知道整個資料的分布情形。
- 若想迅速得知資料的分布情形，可利用第 1 四分位數(Q_1)、中位數(m_e)、第 3 四分位數(Q_3)、最小值(min)、最大值(max)來表示資料的分布，稱為五數綜合(five number summary)，並可畫成盒鬚圖(box-and-whisker plot)，以更清楚得知資料的分布。
- 例子：兩個證券營業員最近 8 個星期的股票交易手續費收入(單位：萬元)如下

鍾樂水	30	63	66	78	82	96	106	270
朱碧霞	64	82	88	90	96	108	128	166

  - 鍾樂水業績的五個數字：$\min=30$ 、 $Q_1=64.5$ 、 $m_e=80$ 、 $Q_3=101$ 、 $\max=270$ 。

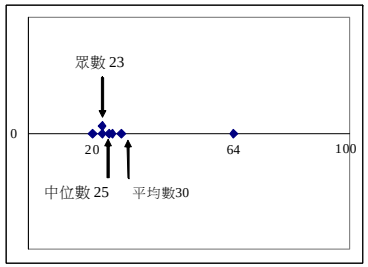
- 鍾樂水業績的盒鬚圖：
  (1) 畫出盒子：以 Q_1 及 Q_3 為盒子的邊界；該盒子顯示了包含的 50%資料的範圍，該範圍稱為四分位距(IQR)。
  (2) 在盒子中畫一條垂直線代表中位數：中位數將盒子中的資料分成兩等分(各 25%)，中位數是資料的中心位置。
  (3) 從盒子的左邊畫一條虛線到最小值，右邊畫一條虛線到最大值；虛線就是所謂的鬍鬚，代表資料的分散情形。



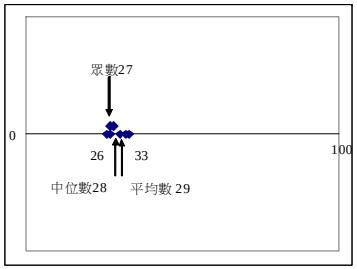
- 盒鬚圖的意義：
  (1) 盒子的寬窄可知居中 50%資料的分散或集中情形：盒子很寬表示居中 50%資料分散度大，反之，則小。
  (2) 盒子的位置顯示資料的偏態：居中表示對稱(左右鬍鬚長度相當)，居左表示右偏(右邊鬍鬚較長)，居右表示左偏(左邊鬍鬚較長)。
- 鍾樂水業績盒鬚圖的解釋：盒子的寬度不大，資料分散程度不大；由盒子位置(鬍鬚相對長短)可知是一個右偏分配。
- 根據朱碧霞業績的五個數字亦可畫出其盒鬚圖：$\min = 64$、$Q_1 = 85$、$m_e = 93$、$Q_3 = 118$、$\max = 166$。
- 兩盒鬚圖比較：朱碧霞業績的盒子較窄，表示其業績較集中；朱碧霞業績的盒子位於鍾樂水的右邊，顯示整體業績狀況較佳；朱碧霞業績的鬍鬚都比鍾樂水短，表示其業績分散程度較小；而且朱碧霞的業績分布也較為對稱。**總結**：朱碧霞的業績比鐘樂水好，且較穩定。

未分組資料分散度的衡量

- 直方圖與盒鬚圖(五數綜合)可看出資料的分散程度，但畢竟只能給我們一個粗略的印象，最好還是用精確的統計測量數來衡量。
- 為何討論資料的分散程度(變異性)是一個重要的問題
  - 當分散程度很小時，資料大多集中於平均數附近(如銀行業的月薪)，則平均數是一個良好的中心位置代表性指標；相反地，若資料的分散度很高(如證券業的月薪)，則平均數就不是一個良好的代表性指標。

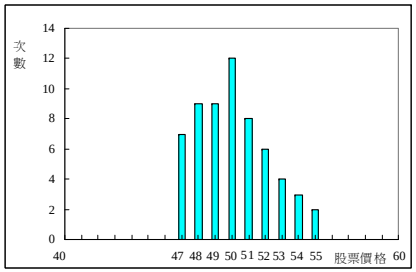


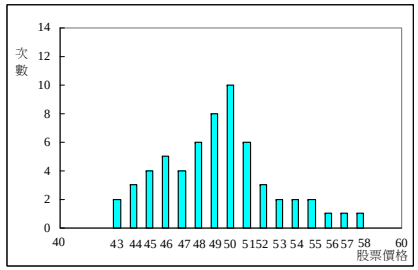
證券業的平均月薪之點圖



銀行業的平均月薪之點圖

- 在比較兩組資料時，不能僅依靠中心位置測量數，亦應考慮資料的分散程度。例如：A(左圖)、B(右圖)兩檔股票價格的次數分配圖；假設 A、B 兩檔股票的平均價格相同(都是 50 元)，但 A 股票的價格分布較集中(集中於 50 元附近)，B 股票價格的變異程度則較大(最高可能上漲至 58 元，最低可能下跌至 43 元)。B 股票的價格較有可能劇烈漲跌，雖然其平均價格跟 A 股票相同，因此對投資人而言，B 股票的投資風險相對較高(大跌的機率較高)。



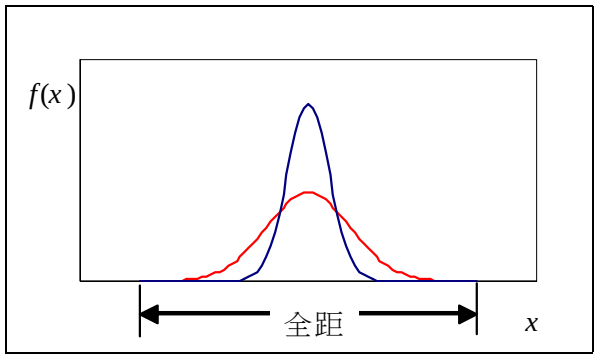


- 分散程度(dispersion)或變異性(variability)的測量數有：全距(range)、四分位距(interquartile range; IQR)、平均絕對離差(mean absolute deviation; MAD)、變異數(variance)、標準差(standard deviation)、變異係數(coefficient of variance)。

全距(Range； R)

- 觀察值中最大值(max)與最小值(min)的距離，以符號 R 表示
  $R = \text{最大值} - \text{最小值} = \max - \min$
  - 全距(R)越大，表示分散程度越大。
  - 例子：銀行業薪資資料的 $\min = 26$、$\max = 33$，因此其全距 $R = 33 - 26 = 7$。證券業薪資的 $\min = 20$、$\max = 64$，全距 $R = 64 - 20 = 44$。由此可知，證券業薪資的差異較大。
- 以全距衡量分散程度的缺點
  - 資料單位不同時不能比較。例如台北市一月的氣溫(度)與七月的降雨量(公厘)，無法根據全距的大小比較其分散程度。

- 就算兩筆資料的單位相同、且全距也相同，也不代表兩筆資料的分散程度一定相同。例如：假設底下兩筆資料的分配具有相同的全距，但其分散程度卻是明顯地不相同。



- 全距只考慮最大與最小兩個觀察值，並未考慮所有的觀察值，故不能精確地反應全體觀察值的分散情形。
- 易受極端值影響。

四分位距(interquartile range; IQR)

- 針對全距易受極端值影響的缺點加以改進。
- 四分位距(IQR)是第 3 四分位數與第 1 四分位數的距離

$$IQR = \text{第3四分位數} - \text{第1四分位數} = Q_3 - Q_1$$

  - 以中間 50%資料的全距(頭尾值的差距)來衡量分散程度。
  - IQR 越大，分散程度越大。
- 例子：在高中男生體重的例子中，$Q_1 = 62$、$Q_3 = 73$，因此 $IQR = Q_3 - Q_1 = 73 - 62 = 11$。

平均絕對離差(mean absolute deviation; MAD)

- 我們偏好以平均數衡量資料的中心位置，主要是因為它使用了所有的觀察值。
  - 在衡量分散度時，最好也將所有觀察值納入考量。
  - 平均絕對離差、變異數、標準差衡量分散度的指標，就使用了所有的觀察值。

- **離均差**(deviation about the mean)、平均數的離差、離差：每一觀察值與平均數間的差距。
  - 若 $x_1,...,x_n$ 等 n 個觀察值是變數 X 的樣本資料，則該樣本的離均差即為[母體資料有類似的定義]

    $x_1 - \bar{X}$ 、 $x_2 - \bar{X}$ 、…、 $x_n - \bar{X}$
  - 小於平均數的觀察值其離均差為負值，大於平均數的觀察值其離均差則為正值。
  - 離均差的絕對值越大，表示該觀察值距離平均數越遠。
  - 若要考慮整體資料的離散情形，應將所有的離均差納入考量，但若使用離均差的平均值並無意義，因為離均差的總和為零 $\sum_{i=1}^{n}(x_i - \bar{X}) = 0$，所以離均差的平均值為零。
  - 較好的方法是僅計算離均差的大小，不考慮正負號，計算所有離均差絕對值的平均值，即平均絕對離差。

- **平均絕對離差**：所有觀察值與平均數之離均差絕對值的平均數，平均絕對離差越大代表資料的分散程度越大。
  - 母體平均絕對離差：若 $x_1,...,x_N$ 等 N 個觀察值是變數 X 的母體資料，則母體平均絕對離差為[μ 為母體平均數]

    $MAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$
  - 樣本平均絕對離差：若 $x_1,...,x_n$ 等 n 個觀察值是變數 X 的樣本資料，則樣本平均絕對離差為[$\bar{X}$ 為樣本平均數]

    $mad = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{X}|$
- 例子：從台北縣新莊市到台北市上班有兩條路可走：一條經三重到台北(縱貫路)，一條經泰山走中山高速公路(中山高)到台北。我們想知道哪條路較好走，分別記錄兩條路線 5 天的行車時間

	A	B	C	D	E	F
1	縱貫路	37	34	39	38	42
2	中山高	44	23	37	31	55

(單位：分鐘)

- 兩條路的行車時間平均都是 38 分鐘，這表示走哪條路都一樣嗎？那可不一定，我們從行車時間分散程度的角度來看。
- 縱貫路的全距為 $42-34=8$ 分鐘，平均絕對離差為 $\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{X}|=\frac{1}{5}\cdot 10=2$ 分鐘；中山高的全距為 $55-23=32$ 分鐘，平均絕對離差為 $\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{X}|=\frac{1}{5}\cdot 46=9.2$。中山高行車時間的差異程度較大；為了準時上班，請問你選哪條路？

縱貫路			中山高		
開車時間	$X-\overline{X}$	$\lvert X-\overline{X}\rvert$	開車時間	$X-\overline{X}$	$\lvert X-\overline{X}\rvert$
37	-1	1	44	6	6
34	-4	4	23	-15	15
39	1	1	37	-1	1
38	0	0	31	-7	7
42	4	4	55	17	17
合計	0	10	合計	0	46

變異數(variance)

- 雖然離差的絕對值是衡量觀察值與平均數距離的最佳方法，但由於絕對值的代數運算較為複雜，因此較難進行統計推論。
- 另一個衡量觀察值與平均數距離的方法，是將離差取平方(同樣可將負離差變成正數)，分散程度則以離差平方的平均數來衡量，這就是平均平方離差(mean squared deviation)或變異數。
- 變異數：觀察值與平均數之離差的平方的平均數
  - 母體變異數：若 $x_1,...,x_N$ 等 N 個觀察值是變數 X 的母體資料，則母體變異數為[μ 為母體平均數]

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

由於 $\sum(x_i-\mu)^2=\sum(x_i^2-2\mu x_i+\mu^2)=\sum x_i^2-2\mu\sum x_i+\sum\mu^2$

$$=\sum x_i^2-2N\mu^2+N\mu^2=\sum x_i^2-N\mu^2$$

故母體變異數又可表達為：$\sigma^2=\frac{1}{N}\left[\sum_{i=1}^{N}x_i^2-N\mu^2\right]$

  - 樣本變異數：若 $x_1,...,x_n$ 等 n 個觀察值是變數 X 的樣本資料，則樣本變異數為[$\overline{X}$ 為樣本平均數]

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

由於 $\sum(x_i-\overline{X})^2=\sum(x_i^2-2\overline{X}x_i+\overline{X}^2)=\sum x_i^2-2\overline{X}\sum x_i+\sum\overline{X}^2$

$$=\sum x_i^2-2n\overline{X}^2+n\overline{X}^2=\sum x_i^2-n\overline{X}^2$$

故樣本變異數又可表達為：$S^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-n\overline{X}^2\right]$

註：樣本變異數公式中，分母是除以 $n-1$，而不是除以 n；這是因為要計算變異數，必須先計算平均數，因此樣本觀察值喪失了一個**自由度**(degree of freedom)。

- **變異數的性質**
  - 變異數的值大於等於 0，若變異數為 0 時，其意義是所有觀察值均相同，沒有變異(分散程度)。

  - 若同一組資料單位不同，其變異數亦不相同。
  - 若資料單位相同時，變異數可作比較。
  - 變異數考慮了每一個觀察數值。
  - 適合代數運算。
  - 適合利用樣本變異數對母體變異數做統計推論
  - 具有複名數(如：元^2)，不易解釋。如電腦價格的變異數的單位為平方元(元^2)，不具意義。

標準差(standard deviation)

- 變異數具有複名數，因此不易解釋。若將變異數開根號，所得到的數字之單位與原始資料單位相同，解決了變異數複名數的缺點
- 變異數開根號所得到的數字稱為標準差。
  - 變異數為『觀察值與平均數的平均平方距離』。
  - 標準差可解釋成『觀察值與平均數的平均距離』。
  - 變異數具有的性質(除了複名數)，標準差均有。

- 標準差之定義：
  - 母體標準差：$\sigma = \sqrt{\sigma^2}$
  - 樣本標準差：$S = \sqrt{S^2}$
- 例子：縱貫路與中山高行車時間的變異數與標準差。
  - 兩條路行車時間樣本觀察值的離均差、離均差平方、離均差平方總和計算如下表：

縱貫路			中山高		
開車時間	$X - \overline{X}$	$(X - \overline{X})^2$	開車時間	$X - \overline{X}$	$(X - \overline{X})^2$
37	-1	1	44	6	36
34	-4	16	23	-15	225
39	1	1	37	-1	1
38	0	0	31	-7	49
42	4	16	55	17	289
合計	0	34	合計	0	600

  - 縱貫路行車時間的變異數為

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{5-1}\cdot 34 = 8.5$$

標準差為

$$S = \sqrt{S^2} = \sqrt{8.5} = 2.92$$

  - 中山高行車時間的變異數為

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{5-1}\cdot 600 = 150$$

標準差

$$S = \sqrt{S^2} = \sqrt{150} = 12.25$$

  - 由標準差的數值可知，中山高行車時間的分散程度大於縱貫路；亦即行車時間的變異程度較大。
- Remark：標準差是以平均數為中心的分散度，當我們以平均數來描述資料的中心位置時，才能以標準差來衡量資料的分散程度
  - 例如：當資料存在離群值時，平均數不具中心位置的代表性，此時以標準差(變異數)衡量分散程度已不可靠。

- 全距、四分位距、平均絕對離差、變異數、標準差，衡量的都是資料的『絕對分散程度』。
  - 若有兩組資料，想要比較這兩組資料的『相對分散程度』時，變異數(標準差)會受到平均數的大小以及衡量單位不同的影響，不能直接用變異數(標準差)來比較兩組資料的分散程度。此時應以**變異係數**為之。

變異係數(coefficient of variation；CV)：Carl Pearson 所提出

- 變異係數：標準差除以平均數，衡量每單位平均數的分散程度

$$\text{變異係數}(CV) = \frac{\text{標準差}}{\text{平均數}}$$

  - 母體變異係數：$CV = \frac{\sigma}{\mu}$
  - 樣本變異係數：$CV = \frac{S}{\overline{X}}$

- 例子：兩基金投資報酬率的差異。假設有兩檔基金過去一段時間報酬率的平均數與標準差如下：

基金類別	平均數(%)	標準差(%)
甲基金	11.32	6.63
乙基金	7.21	4.87

  - 財務金融領域通常用標準差來代表風險程度。
  - 投資人通常喜歡高報酬率，但不喜歡風險。甲基金的平均報酬率較高，但乙基金的風險則較低；此時哪個基金才是較佳的投資標的呢？我們可比較兩基金的變異係數。
  - 甲基金 $CV = 6.63\%/11.32\% = 0.59$
    乙基金 $CV = 4.87\%/7.21\% = 0.68$
  - 此例中，變異係數可解釋成『平均每一單位報酬率所承擔的風險程度』。甲基金每單位報酬率所承擔的風險較低，故甲基金是較佳的投資標的。

柴比氏定理與經驗法則

- 若我們知道某個變數之資料的平均數與標準差，透過柴比氏定理與經驗法則，我們即可得知該變數資料的約略分布情形。
- **柴比氏定理(Chebyshev Theorem)**：
  - 設 X 為一隨機變數，其平均數為 μ，變異數為 σ^2，則對任何正數 $k>1$ 而言

$$P(|X-\mu|<k\sigma)\geq 1-\frac{1}{k^2}$$

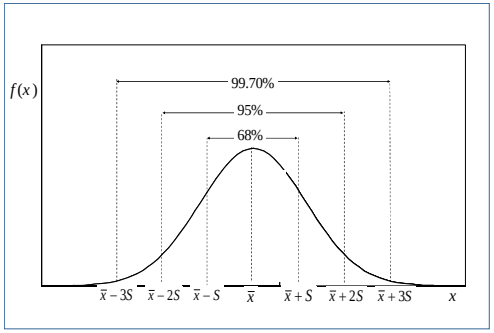
或 $$P(|X-\mu|\geq k\sigma)\leq \frac{1}{k^2}$$

[證明] 留待上到第七章後再證明。

  - 不論資料為何種分配，至少有 $(1-1/k^2)$ 的資料落在距離平均數 k 個標準差的範圍內。k 為大於 1 的任意數，即 $k>1$。

  - 不管資料的分配型態為何，只要知道資料的平均數與變異數，柴比氏定理即可推測資料位於某一範圍內的比率(下限)
  - 若 $k=2$，則至少有 75% $[1-1/k^2=1-1/2^2=3/4]$的觀察值落在距離平均數兩個標準差之內，即 $\bar{X}\pm 2S$ 內 $[\mu\pm 2\sigma]$，或表為 $(\bar{X}-2S,\bar{X}+2S)$ $[(\mu-2\sigma,\mu+2\sigma)]$。若 $k=3$，則至少有 89% $[1-1/k^2=1-1/3^2=8/9]$的觀察值落在 $\bar{X}\pm 3S$ 內 $[\mu\pm 3\sigma]$。若 $k=4$，則至少有 94% $[1-1/k^2=1-1/4^2=15/16]$ 的觀察值落在 $\bar{X}\pm 4S$ 內 $[\mu\pm 4\sigma]$。
  - 例子：某班 100 個學生統計學的平均成績為 75 分，標準差為 5 分。則根據柴比氏定理，統計學成績在 $\bar{X}\pm 2S=75\pm 2\cdot 5=65\sim 85$ 分的同學至少有 75 個；成績在 $\bar{X}\pm 3S=75\pm 3\cdot 5=60\sim 90$ 分的同學至少有 89 個；成績在 $\bar{X}\pm 4S=75\pm 4\cdot 5=55\sim 95$ 分的同學至少有 94 個。

- **經驗法則(empirical rule)**：
  - 日常生活中經常會發現許多資料的分布型態為鐘形(bell-shaped distribution)，此時可用經驗法則判斷資料落於某一範圍內的機率。
  - 若資料為鐘形分配[亦即**常態分配**]，則有 68%的觀察值落在 $\bar{X}\pm S$ 內，有 95%的觀察值落在 $\bar{X}\pm 2S$ 內，有 99.7%的觀察值落在 $\bar{X}\pm 3S$ 內。[與柴比氏定理不同，經驗法則指定分配]



- ***Z* 值(*Z* score)**
  - 若資料的型態為鐘形時，為瞭解某觀察值在資料中的相對位置，可計算該觀察值的 Z 值[將觀察值與平均數的差距表達成標準差的倍數]。
  - Z 值：觀察值減去平均數再除以標準差(或稱**標準化**)

樣本觀察值 x 值的 Z 值：$\frac{x-\bar{X}}{S}$

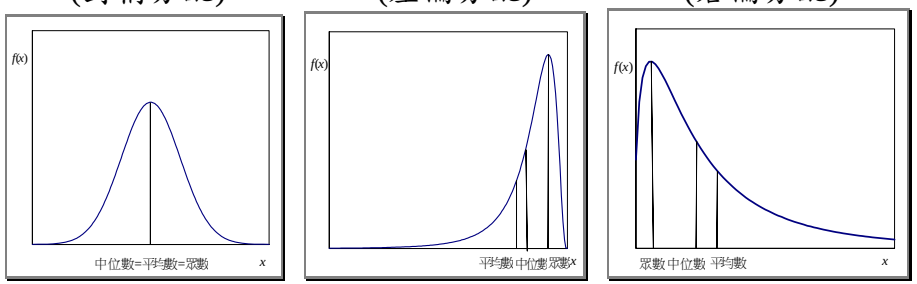
母體觀察值 X 值的 Z 值：$\frac{X-\mu}{\sigma}$

  - 例子：A 班學生的平均成績為 75 分，標準差為 10 分，A 班的甲學生成績為 70 分，則甲學生的 Z 值為 $(70-75)/10=-0.5$；甲學生的分數低於平均數 0.5 個標準差。B 班學生的平均成績為 65 分，標準差為 10 分，B 班的乙學生成績為 70 分，則乙學生的 Z 值為 $(70-65)/10=0.5$；乙學生的分數高於平均數 0.5 個標準差。兩學生的分數同為 70 分，但其分數在班上的地位不同。

未分組資料偏度與峰度的衡量

- 偏態的方向可分為對稱、右偏、左偏三種

(對稱分配)　(左偏分配)　(右偏分配)



  - 對稱：平均數＝中位數＝眾數
  - 左偏：平均數<中位數<眾數
  - 右偏：平均數>中位數>眾數
- 皮爾生(Karl Pearson)偏態(skewness)係數：以平均數與眾數的差距跟標準差的比值來衡量偏態情形。

  - 母體：若取得母體資料，而母體資料的平均數為 μ，標準差為 σ、眾數為 M_0，則皮爾生偏態係數定義為

$$SK_p = \frac{\mu - M_0}{\sigma}$$

  - 樣本：若取得樣本資料，而樣本資料的平均數為 $\bar{X}$，標準差為 S、眾數為 m_0，則皮爾生偏態係數定義為

$$SK_p = \frac{\bar{X} - m_0}{S}$$

  - 偏態情況與皮爾生偏態係數(SK_p)的對應關係
    (1) 對稱：平均數＝眾數，所以 $SK_p = 0$。
    (2) 左偏：平均數<眾數，所以 $SK_p < 0$。
    (3) 右偏：平均數>眾數，所以 $SK_p > 0$。
  - $|SK_p|$越大，表示資料的分布越具偏態
  - 皮爾生偏態係數最大的缺點：若資料的眾數有很多個、或不存在時，無法計算偏態係數。

- 動差法的偏態係數：3 階中央動差與 2 階中央動差 $\frac{3}{2}$ 次方的比值
  - 母體 r 階中央動差(rth central moment)：若 $x_1,...,x_N$ 等 N 個觀察值是變數 X 的母體資料，則母體 r 階中央動差為

$$M_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^r \qquad [\mu \text{ 為母體平均數}]$$

(1) 當 $r=1$，$M_1 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu) = 0$，為母體平均離差。

(2) 當 $r=2$，$M_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sigma^2$，為母體變異數。

(3) 當 $r=3$，$M_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3$，可用來衡量偏態。

(4) 當 $r=4$，$M_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4$，可用來衡量峰度。

  - 樣本 r 階中央動差：若 $x_1,...,x_n$ 等 n 個觀察值是變數 X 的樣本資料，則樣本 r 階中央動差為

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^r \qquad [\bar{X} \text{ 為樣本平均數}]$$

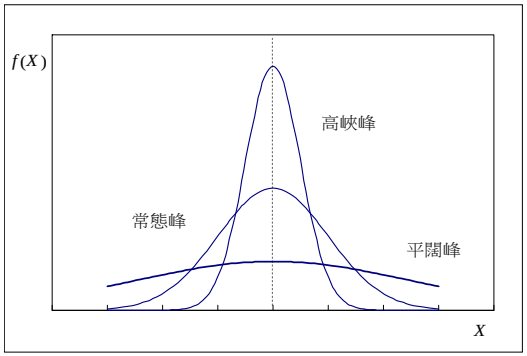
  - 偏態係數：

母體：$\alpha_3 = \frac{M_3}{(M_2)^{3/2}}$

樣本：$\alpha_3 = \frac{m_3}{(m_2)^{3/2}}$

(1) 當 $\alpha_3 = 0$ 時，為對稱分配。
(2) 當 $\alpha_3 > 0$ 時，為右偏分配(正偏分配)。
(3) 當 $\alpha_3 < 0$ 時，為左偏分配(負偏分配)。
(4) 若 $0 \le |\alpha_3| \le 0.5$，則為趨於對稱的分配；若 $0.5 < |\alpha_3| \le 1$，則為稍具偏態的分配；$|\alpha_3| > 1$ 則是極為偏態的分配。

- 峰度(kurtosis)：當資料的分配有集中趨勢時，就會有峰的出現。峰的形態視次數分配集中於平均數、眾數附近的程度，或分散於兩端的情形而定。



- 常態峰(meso kurtosis)：資料的分配呈現一般(正常)形態[峰度與鐘形分配或常態分配一樣]
- 高狹峰(lepto kurtosis)：資料的分布集中於平均數或眾數附近[峰度比鐘形分配或常態分配還大]
- 平闊峰(platy kurtosis)：資料的分布較平均分散於兩端[峰度比鐘形分配或常態分配小]

- 峰度係數：4 階中央動差與 2 階中央動差平方的比值

  母體：$\alpha_4 = \dfrac{M_4}{(M_2)^2}$　　樣本：$\alpha_4 = \dfrac{m_4}{(m_2)^2}$

  (1) 峰度係數一定為正數。

  (2) $\alpha_4 = 3$ 為常態峰；$\alpha_4 > 3$ 為高狹峰；$\alpha_4 < 3$ 為低闊峰。

- **如何檢查極端值（outliers）**：所謂極端值是指與其他大部分的數值比較起來為極小或極大的數值，利用下列步驟可檢查資料是否有極端值。

  步驟 1：將觀察值由小而大排列

  步驟 2：計算出第 1 四分位數 Q_1 與第 3 四分位數 Q_3

  步驟 3：計算四分位距 $IQR = Q_3 - Q_1$

  步驟 4：計算 $Q_1 - 1.5 \times IQR$ 及 $Q_3 + 1.5 \times IQR$

  步驟 5：若觀察值 x 小於 $Q_1 - 1.5 \times IQR$ 或大於 $Q_3 + 1.5 \times IQR$ 則為極端值。

分組資料中心位置的衡量

- 假設已將原始資料經過分組，分組後之資訊如下[令 $\sum_{i=1}^{k} f_i = n$]

組號	組限	組中點(x_i)	次數(f_i)	累加次數(F_i)
1	$L_1 \le x < L_2$	x_1	f_1	F_1
2	$L_2 \le x < L_3$	x_2	f_2	F_2
⋮	⋮	⋮	⋮	⋮
k	$L_k \le x < L_{k+1}$	x_k	f_k	F_k

- **算術平均數**：計算算術平均數必須知道所有觀察值的總和，但分組資料並無法得知每個觀察值的實際數值，因此以每組的組中點取代觀察值，各組總和為組中點與該組次數之乘積。故平均數為

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

- **中位數**：先計算 $\frac{n}{2}$ 或 $\frac{n+1}{2}$ [中位數位置]，確認中位數位於哪一組。假設 L_{m_e} 為中位數(m_e)所在組的組下限，W_{m_e} 為 m_e 所在組的組距，f_{m_e} 為 m_e 所在組的組次數，F_L 為 m_e 前一組的累加次數，則

$$m_e = L_{m_e} + W_{m_e}\left(\frac{\frac{n}{2} - F_L}{f_{m_e}}\right)$$

- **眾數**(粗略法)：找出次數最多的那一組，以該組的組中點為眾數

$$m_0 = \frac{(\text{組上界} + \text{組下界})}{2}$$

- 例子：海之濱營業收入的中心位置。
  - 平均數：根據下表的資料及平均數的公式可知

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{3295}{61} = 54.02$$

	A	B	C	D	E	F	G
1	組號	組限	組距	組中點 x_i	次數 f_i	$f_i x_i$	累加次數
2	1	$20 \le x < 30$	10	25	4	100	4
3	2	$30 \le x < 40$	10	35	7	245	11
4	3	$40 \le x < 50$	10	45	12	540	23
5	4	$50 \le x < 60$	10	55	18	990	41
6	5	$60 \le x < 70$	10	65	11	715	52
7	6	$70 \le x < 80$	10	75	6	450	58
8	7	$80 \le x < 90$	10	85	3	255	61
9					61	3,295	

- 中位數：因 $\frac{n+1}{2}=\frac{61+1}{2}=31$ [n 為奇數]，故中位數位於第 4 組 $(50 \le x < 60)$。因此 $L_{m_e}=50$、$W_{m_e}=10$、$f_{m_e}=18$、$F_L=23$，故中位數為

$$m_e = L_{m_e} + W_{m_e}\left(\frac{\frac{n}{2}-F_L}{f_{m_e}}\right) = 50 + 10 \times \left(\frac{\frac{61}{2}-23}{18}\right) = 54.17$$

- 眾數：次數最多的是第 4 組$(50 \le x < 60)$，因此粗略法眾數為 $m_0=\frac{50+60}{2}=55$。

分組資料等分位置的衡量

- **四分位數**
  - 第 1 四分位數(Q_1)：找出 $\frac{n}{4}$ 位於哪一組；假設 L_{Q_1} 為 Q_1 所在組的組下限，W_{Q_1} 為 Q_1 所在組的組距，f_{Q_1} 為 Q_1 所在組的組次數，F_{Q_1} 為 Q_1 前一組的累加次數，則

$$Q_1 = L_{Q_1} + \frac{\frac{n}{4}-F_{Q_1}}{f_{Q_1}} W_{Q_1}$$

  - 第 2 四分位數(Q_2)：與中位數相同。
  - 第 3 四分位數(Q_3)：找出 $\frac{3n}{4}$ 位於哪一組；假設 L_{Q_3} 為 Q_3 所在組的組下限，W_{Q_3} 為 Q_3 所在組的組距，f_{Q_3} 為 Q_3 所在組的組次數，F_{Q_3} 為 Q_3 前一組的累加次數，則

$$Q_3 = L_{Q_3} + \frac{\frac{3n}{4}-F_{Q_3}}{f_{Q_3}} W_{Q_3}$$

- **十分位數**：找出 $\frac{n \cdot i}{10}$ 位於哪一組(D_i 的位置)；假設 L_{D_i} 為 D_i 所在組的組下限，W_{D_i} 為 D_i 所在組的組距，f_{D_i} 為 D_i 所在組的組次數，F_{D_i} 為 D_i 前一組的累加次數，則第 i 個十分位數 D_i 為

$$D_i = L_{D_i} + \frac{\frac{n \cdot i}{10}-F_{D_i}}{f_{D_i}} W_{D_i}$$

- **百分位數**：找出 $\frac{n \cdot i}{100}$ 位於哪一組(P_i 的位置)；假設 L_{P_i} 為 P_i 所在組的組下限，W_{P_i} 為 P_i 所在組的組距，f_{P_i} 為 P_i 所在組的組次數，F_{P_i} 為 P_i 前一組的累加次數，則第 i 個百分位數 P_i 為

$$P_i = L_{P_i} + \frac{\frac{n \cdot i}{10}-F_{P_i}}{f_{P_i}} W_{P_i}$$

- 例子：某學校大一英文成績的次數分配如下表所示，$n=106$。
  - 第 1 四分位數(Q_1)：$\frac{n}{4}=\frac{106}{4}=26.5$，$Q_1$ 位於第 4 組(60~70)，故 $L_{Q_1}=60$、$W_{Q_1}=10$、$f_{Q_1}=14$、$F_{Q_1}=15$，因此 Q_1 為

$$Q_1 = L_{Q_1} + \frac{\frac{n}{4}-F_{Q_1}}{f_{Q_1}} W_{Q_1} = 60 + \frac{\frac{106}{4}-15}{14} \times 10 = 68.21$$

  - 第 6 十分位數(D_6)：$\frac{n \cdot i}{10}=\frac{106 \cdot 6}{10}=63.6$ 位於第 5 組(70~80)，故 $L_{D_6}=70$、$W_{D_6}=10$、$f_{D_6}=38$、$F_{D_6}=29$，因此 D_6 為

$$D_6 = L_{D_6} + \frac{\frac{n \cdot 6}{10}-F_{D_6}}{f_{D_6}} W_{D_6} = 70 + \frac{\frac{106 \cdot 6}{10}-29}{38} \times 10 = 79.1$$

組號	組限	次數	累加次數
1	30~40	2	2
2	40~50	1	3
3	50~60	12	15
4	60~70	14	29
5	70~80	38	67
6	80~90	33	100
7	90~100	6	106

分組資料分散程度的衡量－變異數與標準差

- 與算術平均數的計算一樣，由於無法得知每個觀察值的實際數值，因此以每組的組中點與平均數的差距取代個別觀察值的離均差，所以每組的離差平方和為 $(x_i-\mu)^2 f_i$ 或 $(x_i-\bar{X})^2 f_i$。
  - 母體變異數與標準差[令 $\sum_{i=1}^{k} f_i = N$]

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i-\mu)^2 f_i \qquad \sigma = \sqrt{\sigma^2}$$

  - 樣本變異數與標準差[令 $\sum_{i=1}^{k} f_i = n$]

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{X})^2 f_i \qquad S = \sqrt{S^2}$$

- 例子：海之濱營業額的變異數與標準差。稍早已算出 $\bar{X}=54.02$，據此可算出每組組中點的離均差；根據下表中的每組離差平方和，可知 $\sum_{i=1}^{k}(x_i-\bar{X})^2 f_i = 13740.98$，故海之濱營業額的變異數為

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{X})^2 f_i = \frac{1}{61-1}\times 13740.98 = 229.016$$

標準差為

$$S = \sqrt{S^2} = \sqrt{229.016} = 15.133$$

	A	B	C	D	E	F
1	組號	組限	組中點 x_i	次數 f_i	$x-\bar{X}$	$(x-\bar{X})^2 f_i$
2	1	$20 \le x < 30$	25	4	-29.02	3,368.64
3	2	$30 \le x < 40$	35	7	-19.02	2,532.32
4	3	$40 \le x < 50$	45	12	-9.02	976.32
5	4	$50 \le x < 60$	55	18	0.98	17.29
6	5	$60 \le x < 70$	65	11	10.98	1,326.16
7	6	$70 \le x < 80$	75	6	20.98	2,640.96
8	7	$80 \le x < 90$	85	3	30.98	2,879.28
9				∑ = 61		13,740.98

另一種分組資訊

- 假設已知每組的次數、平均數、變異數：[假設為樣本資料]

組號	變數符號(X_i)	觀察值個數(n_i)	平均數($\bar{X}_i$)	變異數(S_i^2)
1	X_1	n_1	$\bar{X}_1$	S_1^2
2	X_2	n_2	$\bar{X}_2$	S_2^2
⋮	⋮	⋮	⋮	⋮
k	X_k	n_k	$\bar{X}_k$	S_k^2

則整體資料的算術平均數、變異數為何(X_1、X_2、…、X_k 等變數之資料的整體算術平均數、變異數為何)？

- **算術平均數**：[若為母體資料結果相同]

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i}$$

[證明] 令 x_{ij} 代表變數 X_i 的第 j 個觀察值 $i=1,\cdots,k$，$j=1,\cdots,n_i$，則變數 X_i 的觀察值總和為 $\sum_{j=1}^{n_i} x_{ij}$ $[=n_i\bar{X}_i]$。因此，

$$\bar{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{j=1}^{n_1} x_{ij} + \sum_{j=1}^{n_1} x_{ij} + \cdots + \sum_{j=1}^{n_k} x_{ij}}{\sum_{i=1}^{k} n_i}$$

$$= \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_k\bar{X}_k}{\sum_{i=1}^{k} n_i}$$

$$= \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{\sum_{i=1}^{k} n_i}$$

[簡單證明] 每一組的觀察值總和為 $n_i\bar{X}_i$，故整體資料的總和為 $\sum_{i=1}^{k} n_i\bar{X}_i$，由於全部觀察值個數為 $\sum_{i=1}^{k} n_i$，因此整體的平均數為

$$\bar{X} = \frac{\sum_{i=1}^{k} n_i\bar{X}_i}{\sum_{i=1}^{k} n_i}$$

➢ 特例：若有兩組母體資料，其分別以 X_1、X_2 表示，其觀察值個數分別為 N_1 與 N_2，平均數分別為 μ_1、μ_2，則 X_1 與 X_2 兩組資料之總平均 μ 為：

$$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$$

- **變異數**：

$$S^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2 + \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{\sum_{i=1}^{k} n_i - 1}$$

[證明] 令 x_{ij} 代表變數 X_i 的第 j 個觀察值，$\bar{X}$ 為整體平均數，則根據(樣本)變異數之定義

$$S^2 = \frac{1}{\sum_{i=1}^{k} n_i - 1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{X})^2 = \frac{1}{\sum_{i=1}^{k} n_i - 1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})]^2$$

$$= \frac{1}{\sum_{i=1}^{k} n_i - 1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij} - \bar{X}_i)^2 + 2(x_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) + (\bar{X}_i - \bar{X})^2]$$

$$= \frac{1}{\sum_{i=1}^{k} n_i - 1}\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2\right\}$$

$$= \frac{1}{\sum_{i=1}^{k} n_i - 1}\left\{\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2\right] + \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2\right]\right\}$$

$$= \frac{1}{\sum_{i=1}^{k} n_i - 1}\left[\sum_{i=1}^{k}(n_i - 1)S_i^2 + \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2\right]$$

式中用到

$$\sum_{i=1}^{k}(\bar{X}_i - \bar{X})\left[\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)\right] = \sum_{i=1}^{k}(\bar{X}_i - \bar{X})\cdot 0 = 0$$

因 $\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)$ 為第 i 組離均差之和，故 $\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i) = 0$

➢ 若為母體資料，則整體變異數僅需將 $(n_i - 1)$ 取代成 N_i，將 $(\sum_{i=1}^{k} n_i - 1)$ 取代成 $\sum_{i=1}^{k} N_i$ 即可，亦即

$$\sigma^2 = \frac{\sum_{i=1}^{k} N_i\sigma_i^2 + \sum_{i=1}^{k} N_i(\mu_i - \mu)^2}{\sum_{i=1}^{k} N_i}$$

$$= \frac{\sum_{i=1}^{k} N_i[\sigma_i^2 + (\mu_i - \mu)^2]}{\sum_{i=1}^{k} N_i}$$

式中 N_i 為各組之觀察值個數，σ_i^2 為各組之母體變異數，μ_i 為各組之母體平均數，μ 為整體之母體平均數。

➢ 特例：設 X_1 與 X_2 為兩母體，其平均數與變異數分別為 μ_1、μ_2 及 σ_1^2、σ_2^2，則兩母體的全體變異數為：

$$\sigma^2 = \frac{1}{N_1 + N_2}\{N_1[\sigma_1^2 + (\mu_1 - \mu)^2] + N_2[\sigma_2^2 + (\mu_2 - \mu)^2]\}$$

式中：$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$ (兩母體之全體平均數)