統計學簡介

統計學的意義

- 統計學(statistics)是一種科學方法與原理，其包括資料的蒐集，資料的整理，陳示，分析與解釋，並獲得結論，以幫助做更有效的決策。
- 統計學包含敘述統計學(descriptive statistics)與推論統計學(inferential statistics)。

統計學的意義

完成統計任務的三個步驟爲：

1. 統計資料的搜集與整理
2. 統計資料的分析
3. 統計資料的推論

統計專有名詞

- **母體(population)**
  人們在研究某一現象或問題時，必須針對發生此一現象或問題的對象進行調查研究，**調查研究的全體對象所成之全部集合**即為所謂的母體。
- **樣本(sample)**
  樣本是研究者**從母體中所抽取的部分元素所組成的集合**，亦即為有興趣研究之全體對象的部分集合。

統計專有名詞

- **普查(census)**
  一旦我們確定要研究的對象母體，且母體是有限的，則了解母體最好的方法，就是**對母體內每一個個體加以調查並記錄其特徵**，這種調查方式就稱爲普查(census)。
- **抽樣調查(sampling survey)**
  相對於普查，若**隨機自母體中抽選出一部份具有代表性的個體當作樣本來加以調查**，依據此組樣本進行統計分析，再將所得的結果來推論未知母體，則此種方法即稱爲抽樣調查。

統計專有名詞

- **參數(parameter)**
  爲了推論母體，研究者必須知道描述**母體特徵的某些特徵值**，這些特徵值即稱爲參數或母數。
- **統計量(statistic)**
  統計量是由**樣本中所計算出的量**，其爲**隨機樣本觀察值的函數**，用來推論未知母體參數。

統計專有名詞

- **估計誤差(error of estimation)**
  當我們利用估計量來估計母體參數時，不論用什麼抽樣方式或多精密的測量，**估計量與母體參數間總會有差距**。這樣的差距即稱爲估計誤差。
- **可能誤差(possible error)**
  爲統計數值最末一位數的半個單位。

統計學的發展

劃分階段		代表人物	主要貢獻
古典時期	十六世紀中至十九世紀初	**國勢學派**發展於歐洲大陸，以研究國情爲主 **政治算術學派**發展於英國，舉凡與數字資料有關的社會現象均包含在內	
近代時期	十九世紀初至二十世紀初	卡爾・皮爾生(Karl Pearson,1857－1936)	簡單的統計量，如標準差、相關係數等 迴歸分析的觀念和卡方檢定
		費雪(R.A. Fisher,1890－1962)	實驗設計和隨機過程
現代時期	二十世紀初至今	聶曼(J. Neyman, 1894-1981)及伊根・皮爾生(Egon Pearson)	估計和檢定方面提出理論的基礎
		華德(A. Wald,1902-1950)	逐次分析(sequential analysis)

統計技術

以下列出一些有名的統計檢定方法以及可供驗證實驗數據的程序

- 變異數分析(ANOVA)
- 費雪最小顯著差異法（Fisher's Least Significant Difference test ）
- 學生t檢驗(Student's t-test)
- 曼-惠特尼 U 檢定（Mann-Whitney U）
- 回歸分析（regression analysis）
- 相關性（correlation）
- 皮爾森積矩相關係數(Pearson product-moment correlation coefficient)
- 史匹曼等級相關係數（Spearman's rank correlation coefficient ）
- 卡方分配（chi-square ）

統計的應用

- 政府統計方面 ：如失業率 、人口普查
- 企業統計方面 ：如產品滿意度 、商品佔有率
- 財務金融統計方面 ：如股價指數預測、物價變動
- 教育統計方面 ：如學校經費預估 、師生人數比例
- 農業統計方面 ：如季節變動與農產品收成數量的關係
- 生物統計方面 ：如藥物對疾病的有效度
- 其它統計方面 ：如氣象預測

統計方法的限制

1. 樣本必須具有代表性
2. 必須使用大樣本
3. 必須使用正確資料
4. 態度必須公正客觀
5. 統計數字必須經過比較才有意義
6. 統計結果必須作出合理的解釋

圖解式統計學方法目標

- 圖解式統計學方法具有四個方面的目標
- ⑴ 探究數據集的內容；
- ⑵ 用於發現數據之中的結構；
- ⑶ 檢查統計學模型之中的假設；
- ⑷ 溝通傳達分析結果。
- 如果不採用統計圖形，也就會喪失深入認識數據基礎結構之一個或多個方面的機會。

圖解分析方法的統計學工具

- 這些工具包括散點圖、直方圖、機率圖、殘差圖（residual plot）、箱形圖、塊圖以及雙標圖。探索性數據分析（Exploratory data analysis，EDA）就密切地依賴於這些手段以及與此類似的其他技術方法。
- 圖解分析操作程序不僅僅是在EDA背景下才使用的工具；在檢驗假設、模型選擇、統計模型驗證、估計量（estimator）選擇、關係確定、因素效應判定以及離群值檢出方面，此類圖解分析工具還可以作爲最佳捷徑，用來深入認識數據集。
- 此外，優質的統計圖形還可以作爲一種令人信服的溝通手段，用來向他人傳達存在於數據之中的基本訊息

樣本大小之選擇

- 樣本不要過大，過大浪費成本；但也不要過小，過小則會有太大的抽樣誤差。如何決定適當的樣本大小？在機率抽樣的情況下，有關樣本大小的決定及樣本統計顯著性的判斷，可藉由機率法則的運用。（也就是說，有公式可供計算）
- 但在非機率抽樣的情況下，除了依靠抽樣人員的主觀判斷或假設外，實無客觀之科學方法可資應用。

集中趨勢

- 均數函數應用
  - AVERAGE　傳回引數的平均值
  - AVERAGEA　傳回引數的平均值，包含數字、文字和邏輯值
  - AVEDEV　傳回一組資料與其平均值絕對偏差的平均值
  - HARMEAN　傳回調和平均值
  - COVAR　傳回共變數，即成對偏差乘積的平均數
  - GEOMEAN　傳回幾何平均數
- 中位數應用
  - MEDIAN(number1,[number2], ...)
- 眾數應用
  - MODE(number1,[number2], ...)
- 以線圖表示注重程度差異

中位數

- 中位數（Median）是指將所有數字依大小順序排列後，排列在最中間之數字，其上與其下的數字個數各佔總數的二分之一。也就是說，將所有次數當100%，累積之次數達50%的位置，其觀測值就是中位數（用Me來表示）。
- 於Excel是以MEDIAN()函數來求算中位數，
- 語法：MEDIAN(number1,[number2], ...)
  用以求一陣列或範圍資料的中位數，若這數字爲偶數個數，將計算中間兩個數字的平均值。其算法很簡單，當n爲奇數，按大小排列後，第(n+1)/2個觀測值，就是中位數。當n爲偶數，則取第n/2與(n+2)/2個觀測值之平均數爲中位數。
- 數值1,[數值2],...爲要求中位數之儲存格或範圍引數，最多可達255個。式中，方括號所包圍之內容，表該部份可省略。

- 如：

10, 3, 4, 5, 8, 7, 12

等7個數字資料，n爲7是個奇數，依大小排列後爲：

3, 4, 5, 7, 8, 10, 12

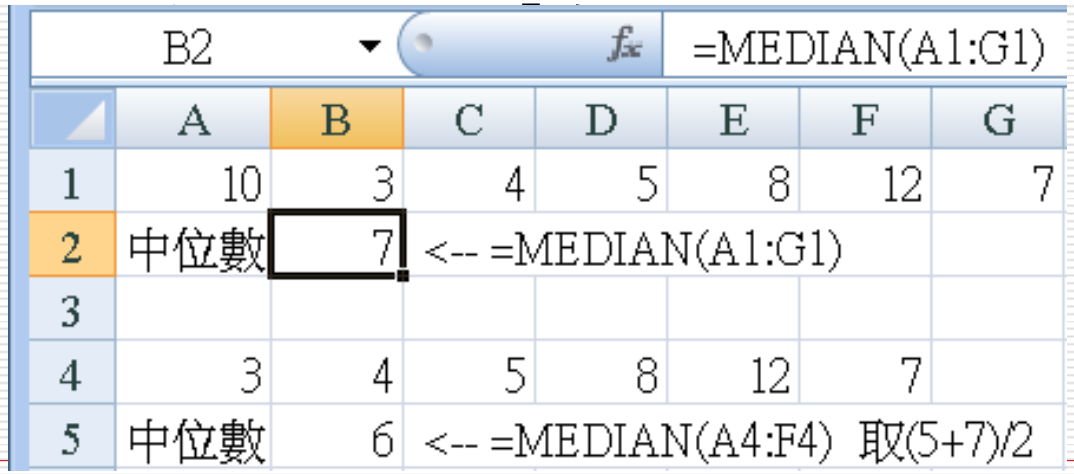
第(7+1)/2=4個觀測值7，就是中位數。而

3, 4, 5, 8, 12, 7

等6個數字資料，n爲6是個奇數，依大小排列後爲：

3, 4, 5, 7, 8, 12

則取第6/2=3與(6+2)/2=4個觀測值之平均數(5+7)/2=6爲中位數。



- 中位數與平均數，均是用來衡量母體的集中趨勢。但中位數不會受極端值影響。如：3, 4, 5, 7, 8, 10, 90

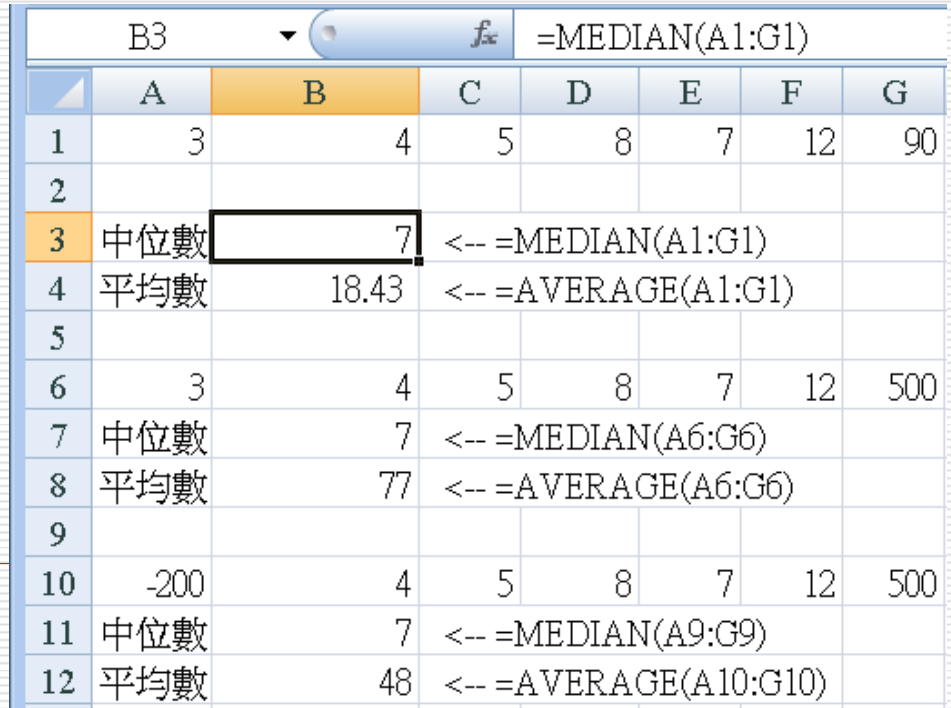
  之平均數爲18.43比六個數字中之五個數字都大，以它來代表這組數字；反不如使用中位數7，來得恰當一點！
- 中位數不會受極端值影響，且無論極端值如何變化，中位數均不變。如：

  3, 4, 5, 7, 8, 10, 500

  或

  -200, 4, 5, 7, 8, 10, 90

  之中位數均還是7



B3 =MEDIAN(A1:G1)

	A	B	C	D	E	F	G
1	3	4	5	8	7	12	90
2							
3	中位數	7	<-- =MEDIAN(A1:G1)				
4	平均數	18.43	<-- =AVERAGE(A1:G1)				
5							
6	3	4	5	8	7	12	500
7	中位數	7	<-- =MEDIAN(A6:G6)				
8	平均數	77	<-- =AVERAGE(A6:G6)				
9							
10	-200	4	5	8	7	12	500
11	中位數	7	<-- =MEDIAN(A9:G9)				
12	平均數	48	<-- =AVERAGE(A10:G10)				

中位數之優點

- 以中位數代表一群數字之集中趨勢的優點爲：
  - 不受極端值的影響
  - 恆爲所有資料的中間分界，它是存在的且易瞭解
  - 對於分配不對稱之資料，中位數比平均數更適合當集中趨勢的代表值。這就是爲何政府機關所公佈之國民所得，常以中位數爲代表值的理由。但對於分配並不是非常不對稱之資料，平均數還是比中位數更適合當集中趨勢的代表值。
- 但其缺點爲：
  - 僅注重中央之數字，忽略了兩端之所有數字
  - 不靈敏，當資料發生變動，中位數並不一定會變動

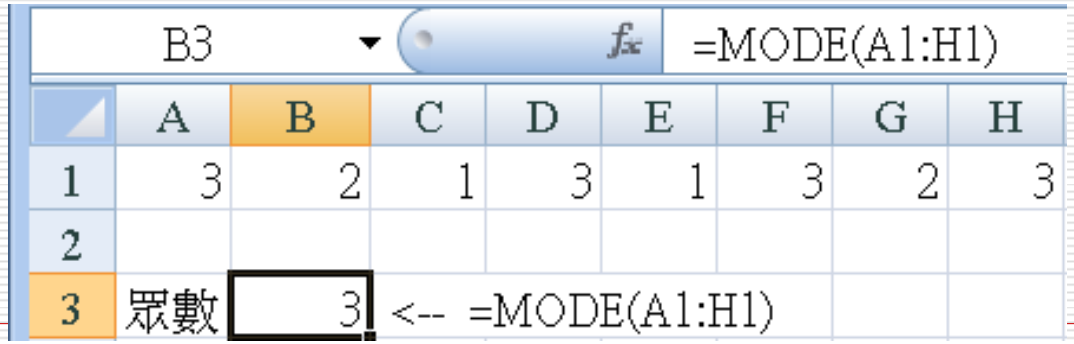
眾數

- 眾數（Mode，以Mo表示）係指在一群體中出現次數最多的那個數值，於Excel係利用MODE()函數來求得。其語法爲：

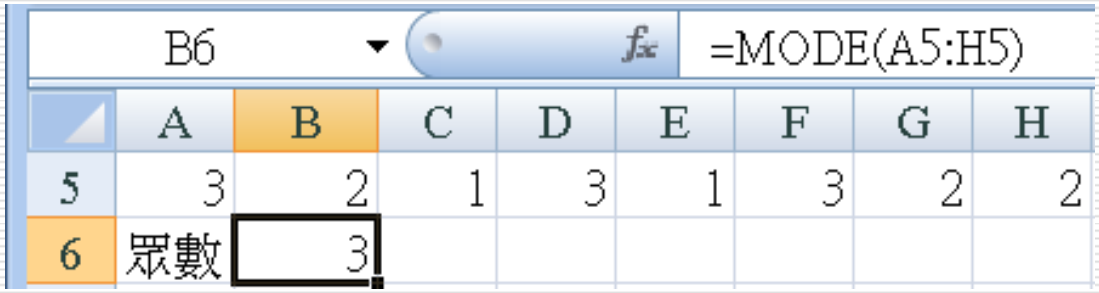
  MODE(數值1,[數值2], ...)

  MODE(number1,[number2], ...)

- 數值1,[數值2], ...爲要求眾數之儲存格或範圍引數，最多可達255個。式中，方括號所包圍之內容，表該部份可省略。如：
  3, 2, 1, 3, 1, 3, 3, 2, 3
  之眾數爲3：



- 眾數、中位數與平均數，均是用來衡量母體的集中趨勢。眾數與中位數是較不會受極端值。不過，眾數並非衡量集中趨勢的好方法，因爲當分配不規則或無顯著之集中趨勢，眾數就無意義。
- 如，可能會同時有好幾個眾數的情況發生：
  3, 2, 1, 3, 1, 3, 2, 2
  之眾數爲3與2，但僅傳回3而已：



- 同時，也可能會沒有眾數！如果資料組中不包含重複的資料點，本函數將傳回#N/A的錯誤值：

眾數之優/缺點

- 眾數之優點爲：
  - 簡單易瞭解
  - 不受兩端極端值影響
- 但其缺點爲：
  - 可能會同時有好幾個眾數的情況發生
  - 也可能會沒有眾數
  - 不靈敏，當資料發生變動眾數並不一定會變動

離散程度

- 均數雖然是一組樣本重要之統計量；但各樣本間之離散程度也是觀察一分配的重要特徵。如果，一分配之離散程度較小，其均數對全體的代表性就較高；反之則否。因此，欲瞭解一分配的基本性質，除需計算均數等集中趨勢數量外；還得衡量其標準差、全距、……等離散程度。
- 最大值減最小值就是全距（range）：
  全距＝最大值－最小值
  全距表示一群體全部數值的變動範圍，是一種離中量數，可用來表示群體中各數字之分散情形，數字大表母體中之數值高的很高，但低的卻很低。

離散程度函數應用

- 四分位差
  - QUARTILE(陣列,類型)
  - QUARTILE(array,quart)
- 百分位數
  - PERCENTILE(陣列,百分比)
  - PERCENTILE(array,percent)
- 變異數
  - 是用來衡量觀測值與平均值間的離散程度，其值越小表母體的離散程度越小，齊質性越高。
  - VARP(number1,[number2],...)
  - VARPA(number1,[number2],...)
- 母體標準差
  - STDEV(number1,[number2],...)
  - STDEVA(number1,[number2],...)
  - STDEVP(number1,[number2],...)
  - STDEVPA(number1,[number2],...)

四分位差

QUARTILE(陣列,類型)

QUARTILE(array,quart)

- 求一個數值陣列或儲存格範圍的第幾個四分位數：將所有數字依大小順序排列後，排列在0%、25%、50% 、75%與100%之數字。如果該位置介於兩數之間，將計算該點左右兩個數字的平均值。
- 陣列是要求得四分位數的數值陣列或儲存格範圍。
- 類型用以指出要傳回的數值：

  0 表最小值(0%處)

  1 表第一個四分位數(25%處)，下四分位數，Q1

  2 表第二個四分位數(50%處)，即中位數，Q2

  3 表第三個四分位數(75%處)，上四分位數，Q3

  4 表最大值(100%處)

百分位數

PERCENTILE(陣列,百分比)

PERCENTILE(array,percent)

- 可用來求一個數值陣列或儲存格範圍的第幾個百分位數：將所有數字依大小順序排列後，排列在百分比所指定位置之數字。如果該位置介於兩數之間，將計算該點左右兩個數字的平均值。
- 陣列是要求得百分位數的數值陣列或儲存格範圍。
- 百分比是介於0～1之百分比數字，如：0.25將求得第一個四分位數(Q_1，25%處，也可以P_{25}表示)，0.5將求得第二個四分位數(Q_2，50%處，也可以P_{50}表示)，即中位數。當其百分比爲10的倍數，則求得者即爲十分位數。如：0.3將求得第三個十分位數D_3（也可以P_{30}表示），0.9將求得第九個十分位數D_9（也可以P_{90}表示）。

母體變異數VARP()與VARPA()

- 母體變異數的計算公式爲：

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$$

即取每一觀測值與其均數間之差異的平方和的算術平均。取其平方就是因爲無論正差或負差，經平方後均爲正值，就不會產生正負相抵銷之情況，以代替取絕對值之麻煩。

- 變異數是用來衡量觀測值與平均值間的離散程度，其值越小表母體的離散程度越小，齊質性越高。於Excel是以VARP()與VARPA()函數來求算母體變異數，其語法爲：
  - VARP(number1,[number2],...)
  - VARPA(number1,[number2],...)
- 式中，方括號包圍之部份表其可省略。數值1,[數值2],...爲要計算變異數之儲存格或範圍引數，它是對應於母群體的1到255個數字引數。

母體標準差STDEVP()與STDEVPA()

- 將母體變異數開根號，即可求得母體標準差。其公式爲：$S=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}}$

  變異數取其平方是因爲要避免正差或負差，產生正負相抵銷之情況。而標準差將其開根號，即是將平方還原，以代替原須取絕對值之麻煩。

- 母體標準差，於Excel也可以STDEVP()與STDEVPA()函數來直接求算。其語法爲：
  - STDEVP(number1,[number2],...)
  - STDEVPA(number1,[number2],...)
- 式中，方括號包圍之部份表其可省略。數值1,[數值2],...爲要計算標準差之儲存格或範圍引數，它是對應於母群體的1到255個數字引數。

變異數與標準差之優缺點

- 變異數與標準差是最常被用來衡量離散程度的方法
- 其優點爲：
  - 感應靈敏
  - 嚴密精確
  - 適於代數處理
  - 受抽樣變動之影響甚小
- 但其缺點爲：
  - 不是簡明易解
  - 計算困難
  - 受極端值影響較大

敘述統計

- 若曾安裝『分析工具箱』，則可以『資料分析』之「敘述統計」增益集，來計算一組資料內之各相關統計值。如：均數、變異數、標準差、全距（範圍）、……等。
- 擬使用『資料分析』之「敘述統計」，來計算運動時間之各敘述統計值。其處理步驟爲：

	A	B
1	編號	每次運動時間/分
2	1	120
3	2	10
4	3	0
5	4	120
6	5	120
7	6	15

參考資源

- 維基百科
- http://zh.wikipedia.org/wiki/%E7%BB%9F%E8%AE%A1%E5%AD%A6