統計學（一）

第二章
(Using Numerical Measures to Describe Data)

授課教師：唐麗英教授

國立交通大學 工業工程與管理學系
聯絡電話：(03)5731896
e-mail：litong@cc.nctu.edu.tw

2013

本課程內容參考書目

- **教科書**
  - P. Newbold, W. L. Carlson and B. Thorne(2007). *Statistics for Business and the Economics*, 7^{th} Edition, Pearson.
- **參考書目**
  - Berenson, M. L., Levine, D. M., and Krehbiel, T. C. (2009). *Basic business statistics: Concepts and applications*, 11^{th} EditionPrentice Hall.
  - Larson, H. J. (1982). *Introduction to probability theory and statistical inference*, 3^{rd} Edition, New York: Wiley.
  - Miller, I., Freund, J. E., and Johnson, R. A. (2000). *Miller and Freund's Probability and statistics for engineers*, 6^{th} Edition, Prentice Hall.
  - Montgomery, D. C., and Runger, G. C. (2011). *Applied statistics and probability for engineers*, 5^{th} Edition, Wiley.
  - Watson, C. J. (1997). *Statistics for management and economics*, 5th Edition. Prentice Hall.
  - 唐麗英、王春和（2013），「從範例學MINITAB統計分析與應用」，博碩文化公司。
  - 唐麗英、王春和（2008），「SPSS 統計分析 」，儒林圖書公司。
  - 唐麗英、王春和（2007），「Excel 統計分析」，第二版，儒林圖書公司。
  - 唐麗英、王春和（2005），「STATISTICA與基礎統計分析」，儒林圖書公司。

如何以量化指標來展示資料

- **連續型資料有以下四個特性：**
  1. Central Tendency (or Location) (集中趨勢)
  2. Dispersion (分散趨勢)
  3. Skewness (偏態)
  4. Kurtosis (峰態)

集中趨勢 Central Tendency

- **集中趨勢**
  - 「集中趨勢指標」是表示一組數據**中央點**位置所在的一個指標。
  - 常用的**集中趨勢指標**
    1. 平均數(mean)
    2. 中位數(median)
    3. 眾數(mode)。

集中趨勢 Central Tendency

- **平均數**
  - 群體平均數：$\mu = \frac{\sum x_i}{N}$
  - 樣本平均數：$\bar{x} = \frac{\sum x_i}{n}$
    - 其中N表群體大小，n表樣本大小。

集中趨勢 Central Tendency

- **例 1：**

  請找出下列群體數據之平均數：0, 7, 3, 9, -2, 4, 6

  [解] μ=3.857

- **例 2：**

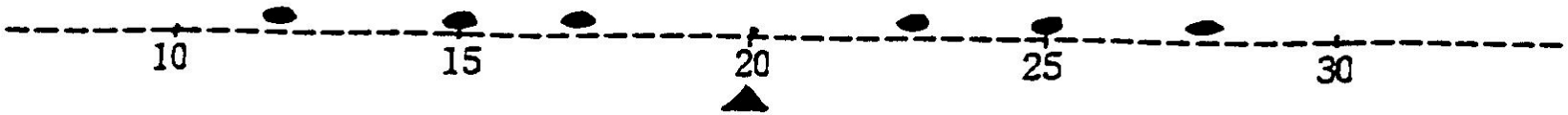
  請找出下列樣本數據之平均數：25, 12, 23, 28, 17 and 15。

  [解] =20

集中趨勢 Central Tendency

- **例 3：**

  將例2之資料繪入下面之點圖中，則平均數為此筆數據之「平衡點」。



集中趨勢 Central Tendency

- **中位數**
  - 將一組數據由小至大排序後，最中間的那一個數值稱為中位數。
  - 群體中位數：η(唸eta)
  - 樣本中位數：$\tilde{x}$
  - 找中位數之方法：

1) 當n=奇數，＝排序第(n+1)/2位之數值。
2) 當n=偶數，＝排序第(n/2)位及第(n/2)+1位的兩數值之平均數。

集中趨勢 Central Tendency

- **例 4：**

  請找出下例樣本數據之中位數：9, 2, 7, 11, 14

- **例 5：**

  請找出下例樣本數據之中位數：9, 2, 7, 11, 14, 6

集中趨勢 Central Tendency

- **眾數**
  - 在一組數據，出現次數最多的數值稱之。

- **例 6：**
  - 請找出下例樣本數據之眾數 3, 3, 2, 1, 4, 2, 3

- **例7：**
  - 請找出下例樣本數據之眾數 3, 2, 1, 1, 4,1, 2,2

集中趨勢 Central Tendency

- **例 8：**

某車商想要生產不同載客人數的車輛，分別為2人座、4人座、5人座、7人座、16人座及20人座的車輛，因此車商調查過去一年的銷售情況，得到結果如下表所示。請問車商應用平均數、中位數或眾數來決定生產何種載客人數的車型最適當？

載客人數	2	4	5	7	16	20
銷售車輛	5	40	46	1	45	43

集中趨勢 Central Tendency

- **何時用平均數？何時用中位數或眾數？**
  - 平均數對離群值非常敏感，而中位數或眾數則對離群值較不敏感。因此，當資料中有離群值時，則使用中位數或眾數，否則，使用平均數。

集中趨勢 Central Tendency

- **例 9：**
  - 1, 3, 4, 6, 6, 9, 13

$\bar{X} = \frac{1+3+\cdots+13}{7} = 6$，中位數＝_____，眾數＝

  - 若在此組數據中有離羣值70：1, 3, 4, 6, 6, 9, 70

則 $\bar{X} = 14.14$，中位數＝_____，眾數＝

集中趨勢 Central Tendency

- **例 10：**

設有A和B兩個大學，其學生每月生活花費如下表（以美金計）：

	A	B
平均數	$3,750	$4,750
中位數	$4,000	$1,250

假如兩個學校的其他因素如：食衣住行等各方面價格均相似，則哪一所學校之學生有較高之月花費？理由為何？

解：_____大學學生每月有較高之生活花費。因其平均數和中位數非常接近，表示A大學約有一半的學生每月花費至少超過該校學生月花費之平均數或中位數。然而，__大學學生之高額月花費指友發生在少數學生身上，約有一半學生之月花費是少於 $1250。

分散趨勢（Dispersion）

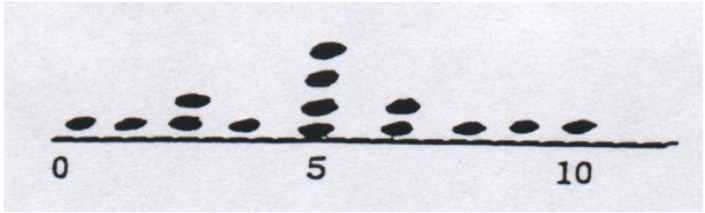
- **分散趨勢**
  - 是表示一組數據間**差異大小**或**數值變化**的一個量數。
  - 四個常用來量測分散趨勢的指標：
    1. **全距（Range）**
    2. **變異數（Variance）**
    3. **標準差（Standard Deviation）**
    4. **變異係數 (Coefficient of Variation, CV)**

分散趨勢（Dispersion）

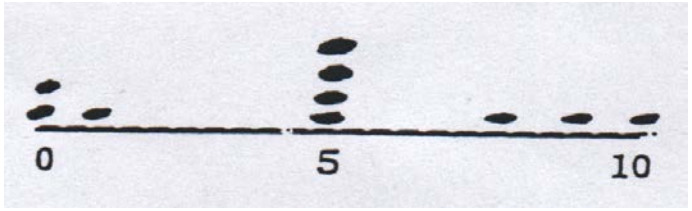
- **例 1：**

以下哪一組資料變異較少？

1)第一組資料:



2)第二組資料:



分散趨勢（Dispersion）

- **全距（R）**
  - 全距是用來衡量一組數據分散程度最簡單的方法
  - 公式： R=最大值－最小值

用全距之缺點：

- 當一組數據中有**離群值**出現或資料**筆數太多**(n>10)時，全距並非一個很好的衡量數據分散程度的量數，因其無法解釋最小值與最大值之間數據分佈的情形。

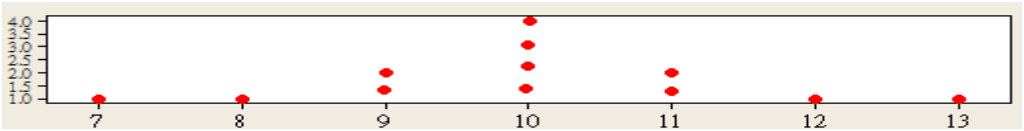
分散趨勢（Dispersion）

- **例 2：**

  以下三組數據有相同之**全距**與**平均**，但有不同之**分佈**。

  – 第一組資料

  – 第二組資料



  – 第三組資料

分散趨勢（Dispersion）

- **變異數和標準差**
  - 群體變異數 $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$
  - 樣本變異數 $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} = \left[\sum_{i=1}^{n} X_i{}^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right] \Big/ (n-1)$
  - 群體標準差 $\sigma = \sqrt{\sigma^2}$
  - 樣本標準差 $S = \sqrt{S^2}$

分散趨勢（Dispersion）

- **例 3：**

  請找出下例樣本數據之平均數、變異數及標準差：

  5, 8, 1, 2, 4

分散趨勢（Dispersion）

- **變異係數 (CV)**
  - 標準差和變異數是衡量一組數據**絕對**變異(Absolute Variation)的指標，即此指標之大小與數據的**單位尺度**有關係，因此，若要比較數組單位尺度不同的數據時，需使用一個衡量**相對變異**的指標，即**變異係數**。

分散趨勢（Dispersion）

- **變異係數 (CV)**
  - 變異係數是一個綜合標準差跟平均數的指標，用來衡量數組資料的相對分散程度。CV指標是以一組數據之標準差佔其平均數的百分比表之，是一個無單位的指標。

換言之，CV代表

群體相對變異：$CV = {}^{\sigma}/_{\mu} * 100\%$

樣本相對變異：$CV = {}^{s}/_{\bar{x}} * 100\%$

分散趨勢（Dispersion）

- **例 4:**

  Given the sample data: 5, 8, 1, 2, 4, find the CV.

Remark: Financial managers sometimes use the CV to measure and compare the riskness of competing portfolios of investments. Those portfolios with higher coefficient of variation go through wider fluctuations in their market value from period than do those portfolios with smaller CV.

分散趨勢（Dispersion）

- **例 5：**

Consider two stocks, A and B. If we take a random sample of the daily closing prices of these stocks, we might find that the respective standard deviations of these closing prices are $S_A = \$0.50$ and $S_B = \$5.00$

According to these standard deviations, we might conclude that the closing prices of stock ____ vary much more than those of the stock ____. Therefore, we might wish to avoid investment in such a volatile(易變的) stock and choose to put our funds into stock ____.

Before we call our broker, however we might be wise to note that

$$\overline{x_A} = \$1.00 \quad \text{and} \quad \overline{x_B} = \$100.00$$

What are the CV of the two stocks? Based upon the CVs for the two stocks, will you put your funds into stock A or stock B ?

分散趨勢（Dispersion）

- **例 6:**

  Two different investments with the stock A mean \$4.00 and the stock B mean \$80.00. Now, the owners are considering purchasing shares of stock A or share of stock B, both listed on the New York Exchange. From the closing prices of both stocks over the last several months the standard deviations were found to be considerably different, with S_A = \$2.00 and S_B = \$8.00. Should stock A be purchased, since the standard deviation of stock B is larger?

偏態（Skewness）

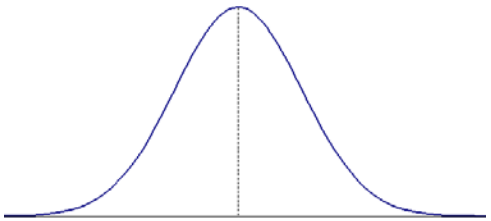
- **偏態（Skewness）**
  - 「偏態」是用來說明一組數據分佈的形態。

- **偏態係數**
  - 樣本偏態係數：$g_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^3 \Big/ (n-1)}{S^3}$
    - 偏態係數＝**0**表示樣本分佈對稱;
    - 偏態係數＝**+**表示樣本分佈偏右;
    - 偏態係數＝**-**表示樣本分佈偏左。

偏態（Skewness）

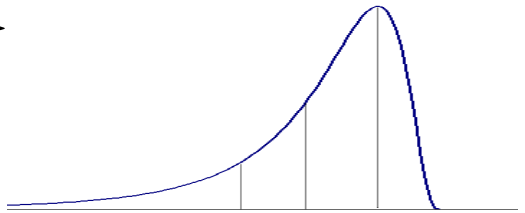
- **單峰分佈有三種形態之偏態：**

1)對稱：平均數____中位數



2)右偏，正偏：平均數____中位數

3)左偏，負偏：平均數____中位數



峰度（Kurtosis）

- **峰度係數：**

  – 樣本峰度係數：$g_2 = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^4 \Big/ (n-1)}{S^3} - 3$

  - 峰度係數＝0：表分佈呈**常態峰**;
  - 峰度係數＜0：表分佈呈**低闊峰**;
  - 峰度係數＞0：表分佈呈**高狹峰**。

峰度（Kurtosis）

- **例7：**

某PC板廠商鍍銅生產線對SEAGATE硬碟PC板的二次銅厚做品檢，抽樣30片量得下列二次銅厚資料，求此批資料之常用統計量並分析之。

單位：μm

123	119	122	121	120
120	121	121	124	118
122	120	125	123	117
121	120	124	122	119
119	118	122	123	121
120	117	119	122	121

峰度（Kurtosis）

- **例7：**
  - 【Excel 報表】

二次銅厚	
平均數	120.8
標準誤	0.372627
中間值	121
眾數	121
標準差	2.04096
變異數	4.165517
峰度	-0.43902
偏態	0.003129
範圍	8
最小值	117
最大值	125
總和	3624
個數	30
信賴度(95.0%)	0.762107

非中趨勢指標(Measures of Noncentral Tendency)

- **非中趨勢指標**
  1. 百分位Percentiles
  2. 四分位數Quartiles

- **如何找各百分位數及四分位數?**
  - **將資料先排序!**

非中趨勢指標(Measures of Noncentral Tendency)

- **百分位數**
  - 第p百分位數代表在數據資料中有p*100%的資料小於或等於此第p百分位數。

- **四分位數**
  - 第一四分位數(The first quartile): Q1 = 25th Percentile
  - 第二四分位數(The second quartile): Q2 = 50th Percentile
  - 第三四分位數(The third quartile): Q3 = 75th Percentile

非中趨勢指標(Measures of Noncentral Tendency)

- **例 8：**

  Find the first and third quartiles of the ten accounting final exam grades listed here from lowest to highest grade :

  51 63 65 70 73 77 79 79 85 88

非中趨勢指標(Measures of Noncentral Tendency)

- **例 9：**

  Find Q_1 and Q_3 for the following data:

  128.3 116.6 132.1 93.9 125.0

  106.5 152.4 105.8 136.7

數據的應用

- 如何決定數據分布之情形?

1. The Empirical Rule(經驗法則)
2. The Chebyshev's Rule(柴比雪夫法則)

經驗法則（The Empirical Rule）

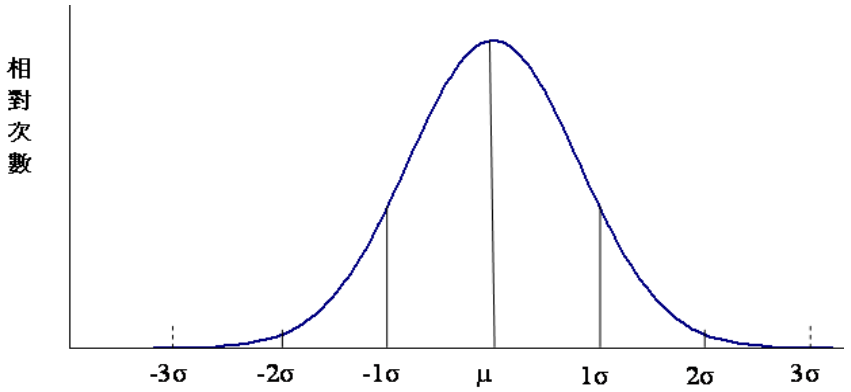
- **經驗法則（The Empirical Rule）**
  又稱68%-95%-99.73%法則

  - 若數據資料呈**鐘形**分布，則約有

    68.26%的數據在μ±σ 範圍內

    95.44%的數據在μ±2σ範圍內

    99.73%的數據在μ±3σ範圍內



經驗法則（The Empirical Rule）

- **例 10：**

  某工程師想要研究某種蝕刻液(etching solution) 對晶圓蝕刻率之影響。該工程師隨機挑選了40片晶圓，得到平均蝕刻率12.8%與標準差1.7%。請以經驗法則來描述此樣本資料

  解：n=40 , =12.8 , S=1.7

  - 大約有68.26%晶圓的蝕刻率是介於12.8±1.7=(11.1,14.5)%。
  - 大約有95.44%晶圓的蝕刻率是介於12.8±2*1.7=(9.4,16.2)%。
  - 大約有99.73%晶圓的蝕刻率是介於12.8±3*1.7=(7.7,17.9)%。

柴比雪夫法則 (The Chebyshev's Rule)

- **柴比雪夫法則 (The Chebyshev's Rule)**
  - 不論數據呈何種分布，至少有$(1-{}^{1}/_{K^2})*100\%$ 的數據會落在μ±kσ範圍內。

  - 例：

至少有$(1-{}^{1}/_{1^2})*100\%=0\%$ 的數據在μ±1σ 範圍內

至少有$(1-{}^{1}/_{1.5^2})*100\%=55.6\%$ 的數據在μ±1.5σ 範圍內

至少有$(1-{}^{1}/_{2^2})*100\%=75\%$ 的數據在μ±2σ 範圍內

至少有$(1-{}^{1}/_{3^2})*100\%=89\%$ 的數據在μ±3σ 範圍內

柴比雪夫法則 (The Chebyshev's Rule)

- **例 11：**

  Trucks traveling on 國道一號weigh an average of 12.5 tons and have a standard deviation of 2.2 tons. Describe the data using the Chebyshev's Rule.

Measures of Relationship between Variables

Measures of Relationship between Variables

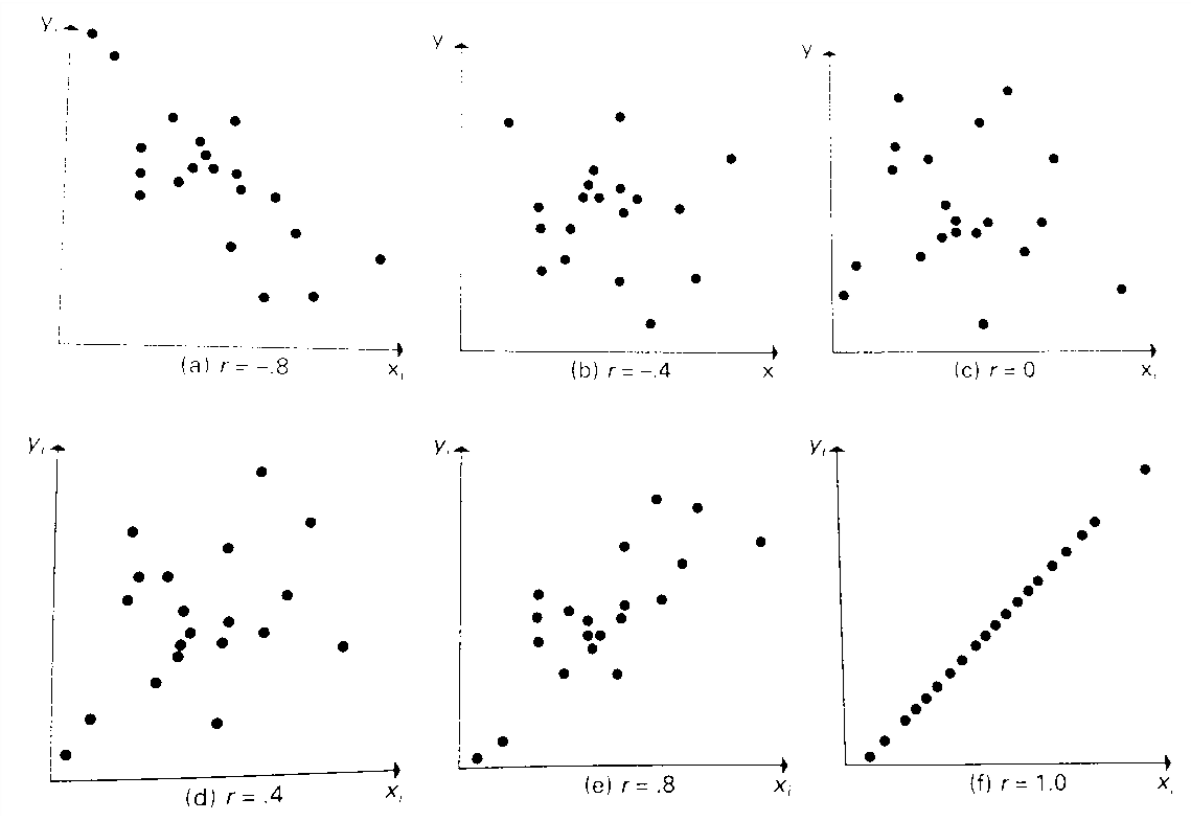
- **Covariance(共變異數)**
  - Covariance (Cov) is a measure of the linear relationship between two variables.
  - Population covariance $\mathrm{Cov}(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$
  - Sample covariance $\mathrm{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

Measures of Relationship between Variables

- **Correlation Coefficient(相關係數)**
  - The Correlation Coefficient is computed by dividing the covariance by the product of the standard deviations of the two variables.
  - Population Correlation Coefficient $\rho = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$
  - Sample Correlation Coefficient $r = \frac{\mathrm{Cov}(x,y)}{s_x s_y}$

Measures of Relationship between Variables

- **Correlation Coefficient(相關係數)**



Measures of Relationship between Variables

- **例 1：**

A corporation administers an aptitude test to all new sales representatives. Management is interest in the extent to which this test is able to predict weekly sales of new representatives. Table records aptitude test scores for a random sample of eight representatives.

Test Score, x	12	30	15	24	14	18	28	26	19	27
Weekly Sales, y	20	60	27	50	21	30	61	54	32	57

Measures of Relationship between Variables

- **例 1：**

[解] $\bar{x} = 21.3,\ \bar{y} = 41.2$

x	12	30	15	24	14	18	28	26	19	27	$\sum$ x=
y	20	60	27	50	21	30	61	54	32	57	$\sum$ y=
xy		1800	405	1200	294	540	1708	1404	608	1539	$\sum$ x y=

$$\mathrm{Cov}(x, y) = s_{xy} =$$

$$r = \frac{\mathrm{Cov}(x,y)}{s_x s_y} = \frac{106.93}{\sqrt{42.01}\sqrt{278.4}} = 0.989$$

本單元結束