國立中大壢中
第二冊數學講義

第四章　數據分析

4-1 一維數據分析

4-2 二維數據分析

班級：_____　　座號：_____

姓名：_____

好棒個數：_____

簽名：_____

4-1 分析一維資料

次數分配表

1.資料的分類：

(1)離散型資料：

資料能夠依照類別分組統計其個數者。

・次序資料：如教育程度及比賽的名次等資料的分類。

・名目資料：如宗教信仰及性別等資料的分類。

(2)連續型資料：

資料不能夠依照類別分組計數，而必須利用數線上的區間來分組統計其個數者

EX：如身高（150 公分～180 公分），體重（50 公斤～80 公斤），年齡（15 歲～60 歲）等

2.次數分配表：

次數分配表是將一群資料分成"幾個組"，然後再將每一組中資料出現的次數，作成一個表，它可更進一步作為畫圖的依據。

(1)離散型資料的次數分配表：

依照資料中各類別分組，再統計其個數，製成次數分配表。

(2) 連續型資料的次數分配表：

【步驟】求 全 距：資料中最大數據而最小數據的差，稱為全距。

定 組 數：將統計資料分類稱為分組，分組的個數稱為組數。
組數的多寡視實際需要而定，一般大約在 5~25 組左右。

定 組 距：每一分組區間的長度稱為該組的組距，一般皆採用每分組的組距都相等的方式分組，故其組距=全距÷組數。

定 組 限：每一組距中最大的數值與最小的數值，分別稱為該組的上限與下限，並統稱該組的組限(規則：含下限但不含上限)

歸組劃記：將每一個資料在對應的組內劃記，五劃為一束（常記作𝍸或正），以便計算。

計算次數：劃記完畢後，計算各組中記號的次數，填入次數欄，並核對次數的總和與資料個數是否相等。

3.相對次數分配表：

將次數分配表中各組中的次數，轉為佔總資料的百分比。

4.累積次數分配表：[適用於連續型資料]

各組別由小而大依序排列，統計比上界小的資料個數。

5.相對累積次數分配表：[適用於連續型資料]

將累積次數分配表中各組中的累積次數，轉為佔總資料的百分比。

例 1：中壢高中高二 30 位學生球類偏好：[離散型資料]

1	2	3	4	5	6	7	8	9	0
籃球	排球	籃球	合球	籃球	羽球	羽球	籃球	籃球	籃球
11	12	13	14	15	16	17	18	19	20
羽球	籃球	籃球	排球	籃球	排球	籃球	籃球	排球	排球
21	22	23	24	25	26	27	28	29	30
羽球	羽球	籃球	羽球	羽球	合球	籃球	合球	籃球	籃球

將上述資料，製作其次數分配表、相對次數分配表。

例 2：某班 50 位學生，身高分別記錄如下：(單位：公分) [連續型資料]

172.2	151.6	132	153	157.9	137	147	146.4	161.2	145
154.6	158	174.3	154.7	135	164	136	167.1	147.4	166.2
127.4	150.2	181.3	155.1	153.3	159	178.2	149	158.2	178.2
138.1	142.4	161.6	140	169.2	166.5	154.6	146	153	129.3
150.5	157.6	144.9	155	144.2	148.4	166.8	163	147.7	168.3

試依 120～130、130～140、140～150、150～160、160～170、170～180、180～190，將上述資料分成 7 組 (各組含下界、不含上界)，製作其次數分配表、相對次數分配表、累積次數分配表、相對累積次數分配表

統計圖

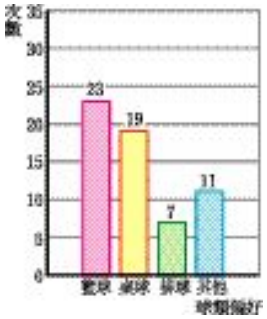
1.離散型資料圖：

(1) 長條圖：

每類用分離的長條表示，長方形的高度表示次數。橫坐標為類別，縱坐標為次數。

(2)圓面積圖：

每類別用同一個圓中的扇形表示，次數與面積成正比。



【長條圖】

【圓面積圖】

2.連續型資料圖：

(1)直方圖：

用連接的長方形表示每一區間，各組以組距為寬，次數為高。橫坐標為數值區間，縱坐標為次數。

(2)次數分配折線圖：

每組資料的(組中點, 次數)為座標，連成的折線所組成。左、右兩端各增加一組，其對應高度都是 0。

(3)相對次數分配折線圖：

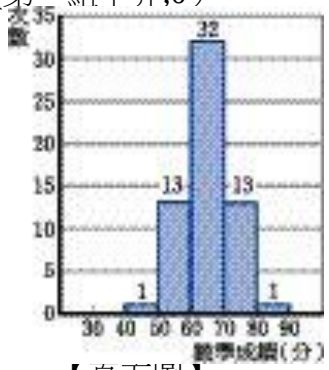
每組資料的(組中點, 相對次數)為座標，連成的折線所組成。左、右兩端各增加一組，其對應高度都是 0。

(4)累積次數分配曲線圖：

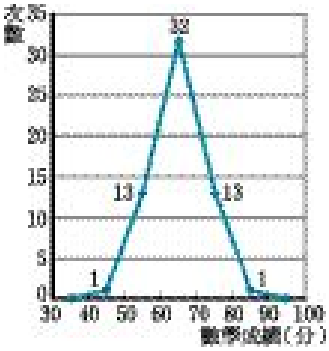
每組資料的(上界,累積次數)為座標，連成的折線所組成。最左邊連到(第一組下界,0）

(5)相對累積次數分配曲線圖：

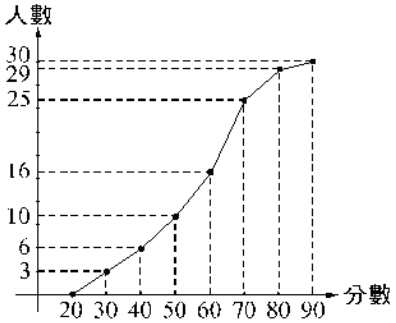
每組資料的(上界,相對累積次數)為座標，連成的折線所組成。最左邊連到(第一組下界,0）



【直方圖】

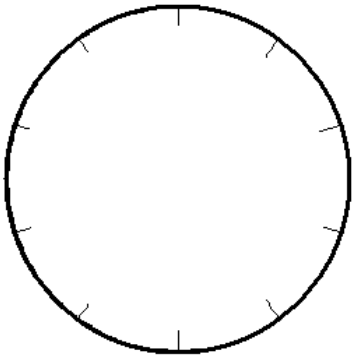


【次數分配折線圖】



【累積次數分配曲線圖】

例 3：製作例 1 資料中的長條圖、圓面積圖



例 4：製作例 2 資料中的直方圖、次數分配折線圖、累積次數分配曲線圖

統計圖描述

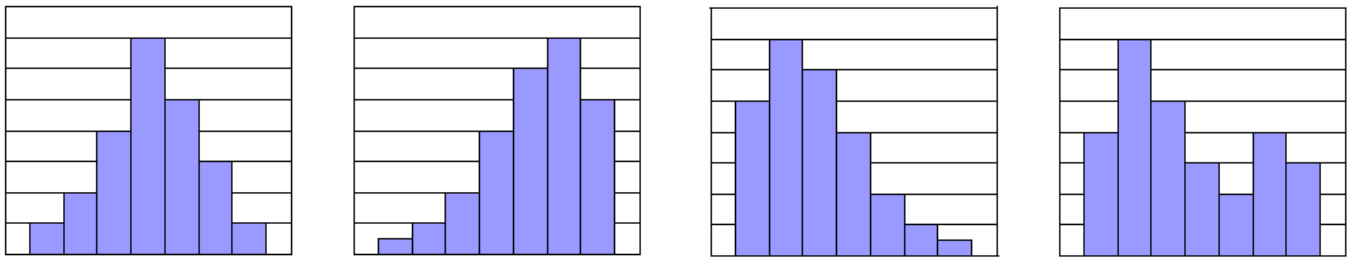
1. 統計圖描述

(1)單峰：直方圖或次數分配折線圖，中間高兩邊低、對稱且只有一個高峰。

(2)左偏：直方圖或次數分配折線圖，左邊尾巴拉的很長。

(3)右偏：直方圖或次數分配折線圖，左邊尾巴拉的很長。

(4)雙峰：直方圖有兩個高峰。EX:可能發生在資料是男女合併的身高資料。



【單峰】　【左偏】　【右偏】　【雙峰】

表達集中趨勢的統計量

1.統計量：

就是由一組樣本資料所算出的單一數值。

2.平均數(*mean*)的意義：

在統計學上，常用一個**平均數**來表示母群體的集中趨勢，做為統計分析衡量標準。

常用的平均數有算數平均數、加權平均數、中位數、眾數。

(算術)平均數

1. (算術)平均數 (*arithmetic mean*)：以 μ 、 $\overline{X}$ 表之

設 N 個母體 $x_1, x_2, \ldots, x_N$ ，抽出 n 個樣本 $x_1, x_2, \ldots, x_n$ ，則

(1)未分組資料求算術平均數

定義：母體算術平均數 $\mu = \dfrac{x_1 + x_2 + \ldots + x_N}{N} = \dfrac{1}{N}\sum_{i=1}^{N} x_i$ 。

樣本算術平均數 $\overline{X} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ 。

- 希望能用 $\overline{X}$ 準確的推估 μ 。

	組中點	次數	次數
$L_1 \sim U_1$	x_1	f_1	f_1
$L_2 \sim U_2$	x_2	f_2	f_2
⋮	⋮	⋮	⋮
$L_k \sim U_k$	x_k	f_k	f_k
總計		N	n

(2)已分組求算術平均數：

設 N 個母體及 n 個樣本的資料次數分配表如下，

定義：母體算術平均數 $\mu = \dfrac{f_1 \cdot x_1 + f_2 \cdot x_2 + \ldots f_k \cdot x_k}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i \cdot x_i$ 。

樣本算術平均數 $\overline{X} = \dfrac{f_1 \cdot x_1 + f_2 \cdot x_2 + \ldots f_k \cdot x_k}{n} = \dfrac{1}{n}\sum_{i=1}^{k} f_i \cdot x_i$ 。

2.算術平均數的特性：

(1). $\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{X}\right)=0$ 。　　(2). $\sum_{i=1}^{n}x_i=n\overline{X}$ 。

(3).公式簡單，適於代數方法計算，故在統計學上最常使用。

(4).因算術平均數視每一個資料皆具有相同的重要性，故其易受極端值影響。

3.資料的平移及伸縮對算術平均數的影響：

若資料 $x_1,x_2,\ldots,x_n$ 的算術平均數為 $\overline{X}$ ，$i=1,2,3,\ldots,n$ ，則

(1). x_i+A 的算術平均數為_____。

(2). Bx_i 的算術平均數為_____。

(3). Bx_i+A 的算術平均數為_____。

例 5：求 10 個數 1981，1988，2000，1892，1997，2101，2014，2022，1996，2019 的算術平均數。【2001】

例 6：某班 50 名學生在第一次段考的數學成績之次數分配表如下，求算術平均數。【70.4】

成績	人數 f_i	組中點 x_i	$fi\cdot x_i$	$y_i=\frac{x_i-A}{B}$	$y_i f_i$
30–40	1				
40–50	2				
50–60	8				
60–70	12				
70–80	15				
80–90	9				
90–100	3				
total	50				

1. 加權平均數：(*weighted mean*)

設 n 個數值資料 $x_1, x_2, x_3, \ldots, x_n$ 的權數分別為 $w_1, w_2, w_3, \ldots, w_n$ ，

定義：加權平均數 $W = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ 。

2.加權平均數的特性：

若各個資料的重要性不同時，賦予各項數值不同的權數(輕重)，則使用加平均數

例 7：某生各科成績如下,試求其加權平均數。 【80.25】

科目	每週上課時數	成績
國文	6	80
英文	6	90
數學	6	76
物理	3	72
化學	3	78

中位數

1.中位數(*median*) ：常以符號 Me 表之。

(1)未分組求中位數：

設 n 個數值資料由小而大排列如下： $x_1 \le x_2 \le x_3 \le \ldots \le x_n$

定義：若 n 為奇數，則中位數 $Me = x_{\frac{n+1}{2}}$ 。

若 n 為偶數，則中位數 $Me = \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ 。

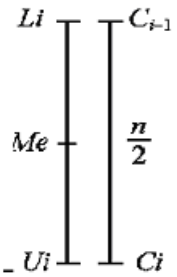
(2)已分組求中位數：

假設在已分組的情況時， Me 發生在第 $\dfrac{n}{2}$ 個數，將 n 個數值資料分組整理，

得累積次數分配表如下，若 $C_{i-1} < \dfrac{n}{2} < C_i$ ，則 Me 落在 $Li \sim Ui$ 這一組，即 $Li < Me < Ui$ 。

假設各組中，各數值均勻的分佈，則利用內插法可得 $\dfrac{Me - Li}{Ui - Li} = \dfrac{\frac{n}{2} - C_{i-1}}{C_i - C_{i-1}}$ ，

	累積次數
$L_1 \sim U_1$	C_1
$L_2 \sim U_2$	C_2
⋮	⋮ ⋮
$Li \sim Ui$	C_i
$L_k \sim U_k$	C_k



2.中位數的特性：不易受極端值影響。

3. 資料的平移及伸縮對中位數的影響：

設 n 個數值資料 $x_1, x_2, x_3, \ldots, x_n$ 的中位數為 Me ，設 $A, B \in R$ ， $i = 1,2,3,\ldots,n$ ，則

(1). $x_i + A$ 的中位數為_____。

(2). Bx_i 的中位數為_____。

(3). $Bx_i + A$ 的中位數為_____。

例 8：設甲、乙兩組第一次月考數學分數如下。

甲組：57,49,40,60,75,45,46,50,53,70,55,31,58

乙組：72,69,50,80,67,76,83,53,68,60

（1）甲組的中位數為_____（2）乙組的中位數為_____ 【（1）53（2）68.5】

例 9：中壢高中高二某班某次段考數學成績如下，

試求：中位數（四捨五入至整數） 【65】

分數	人數
20~30	1
30~40	2
40~50	7
50~60	9
60~70	13
70~80	10
80~90	6
90~100	2

眾數

1.眾數(*mode*)：常以符號 M_0 表之。

資料中出現次數最多的資料值，稱為眾數。

(1)未分組求眾數：將資料排序後，計算出現次數最多的數值即為眾數。

(2)已分組求眾數：從次數分配直方圖觀察，在最高峰值那一組，即為眾數組。

2. 資料的平移及伸縮對眾數的影響：

設 n 個數值資料 $x_1, x_2, x_3, \ldots, x_n$ 的眾數為 Mo ，設 $A, B \in R$ ， $i = 1,2,3,\ldots,n$ ，則

(1). $x_i + A$ 的眾數為_____。

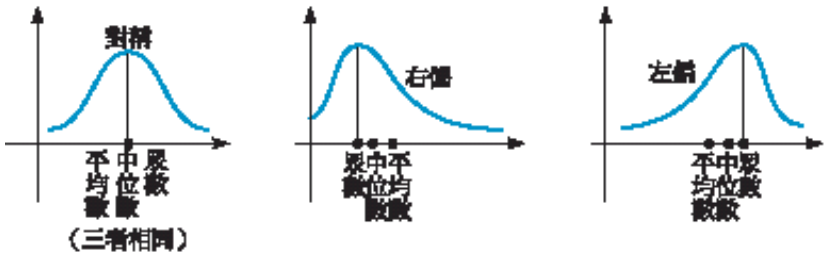
(2). Bx_i 的眾數為_____。

(3). $Bx_i + A$ 的眾數為_____。

例 10：25 個學生舉辦投籃比賽,每個學生投籃 10 次,進球 7,4,6,2,2,7,4,4,3,1,5,4,5,9,4,1,4,9,8,5,4,4,1,4,6,求這群學生投籃進球數的眾數。【4】

眾數、中位數、平均數

1. 對稱、右偏、左偏時，統計量相對位置



表達離散趨勢的統計量

1. 表達離散趨勢的統計量：

表達資料分散狀況的量測，也就是量測資料離中心點多遠的指標。本節介紹的離散趨勢統計量有全距、四分位距、平均絕對離差、變異數

全距

1.全距(*range*)：常以 R 表示

(1)未分組資料：定義全距 $R = x_{\max} - x_{\min}$

(2)已分組資料：定義全距以最大組的上限 U_k 代表最大數，最小組的下限 L_i 代表最小數，故全距 $R = U_k - L_i$ 。

2. 全距的特點：

(1) 計算容易，意義簡明。
(2) 容易受極端值影響，不能顯出分布的情形。

3. 資料的平移及伸縮對全距的影響：
設 n 個數值資料 $x_1, x_2, x_3, \ldots, x_n$ 的全距為 R，設 $A, B \in \mathbb{R}$，$i = 1,2,3,\ldots,n$，則
(1). $x_i + A$ 的全距為_____。
(2). Bx_i 的全距為_____。
(3). $Bx_i + A$ 的全距為_____。

例 11：中壢高中某班 38 位學生參加數學考試，其分數如下：
90 , 97 , 59 , 95 , 78 , 70 , 95 , 71 , 69 , 44 , 80 , 75 求其全距。 【53】

例 12：中壢高中某班 50 位學生參加數學考試，學生的成績整理如下 (原始資料已捨棄不用)，
求其全距。 【60】

組別(分)	30 ~ 40	40 ~ 50	50 ~ 60	60 ~ 70	70 ~ 80	80 ~ 90
人數(人)	4	6	8	14	10	6

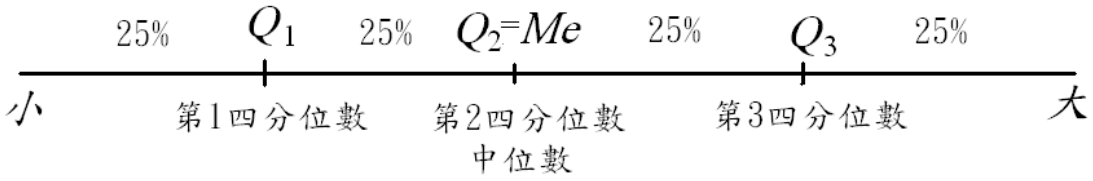
四分位距

1.四分位距(*inter-quartile range*)：簡寫成 *IQR*

四分位距為以資料居中 50%的數值資料所分散距離的差量。

(即第 25%的數值資料與第 75%的數值資料的差)

2.第 k 四分位數 Q_k（k＝1，2，3）的求法：



(1)未分組資料：步驟一、將 n 個資料依小到大排序 $x_1 \le x_2 \le x_3 \le \ldots \le x_n$。

步驟二、計算位置，令 $i = \frac{k}{4} \times n$（k＝1，2，3）

步驟三、定義：若 i 不是整數，則 $Q_k = x_{[i]+1}$

若 i 為整數，則 $Q_k = \frac{x_i + x_{i+1}}{2}$ 。

(2)已分組資料：求第 1 四分位數 Q_1 ，第 3 四分位數 Q_3 ，方法和求中位數 Me 類似

3.四分位距：

第 3 四分位數與第 1 四分位數的差距稱為四分位距。即 $IQR = Q_3 - Q_1$。

・四分位**距** = 四分位**差**。

4.四分位距特性：

一組資料的第 k 四分位數 Q_k ($k = 1, 2, 3$)至少有$\frac{k}{4}$的資料小於或等於 Q_k，也至少有($1 - \frac{k}{4}$)的資料大於或等於 Q_k。

5.資料的平移及伸縮對四分位距的影響：

設 n 個數值資料 $x_1, x_2, x_3, \ldots, x_n$ 的四分位距為 IQR，設 $A, B \in \mathbb{R}$ ，$i = 1,2,3,\ldots,n$ ，則

(1). $x_i + A$ 的四分位距為_____。

(2). Bx_i 的四分位距為_____。

(3). $Bx_i + A$ 的四分位距為_____。

・討論：學科能力測驗成績的 5 標如下，前標、均標、後標分別等於哪個統計量？

頂標：成績位於第 88 百分位數之考生級分。

前標：成績位於第 75 百分位數之考生級分。

均標：成績位於第 50 百分位數之考生級分。

後標：成績位於第 25 百分位數之考生級分。

底標：成績位於第 12 百分位數之考生級分。

例 13：分別求下列各資料的中位數、第 1 四分位數、第 3 四分位數、四分位差

（1）6 個數值資料：32 , 39 , 42 , 58 , 69 , 82

（2）8 個數值資料：25 , 31 , 32 , 39 , 42 , 58 , 69 , 82

例 14：求下列資料第 1 四分位數Q_1、第 3 四分位數Q_3、四分位差$Q.D.$，由$Q.D.$說明此資料的集中趨勢。【$Q_1 = 53.125$、$Q_3 = 75.5$、$Q.D. = 22.375$，全班有一半的學生（25 人），其成績介於 53.125 ~ 75.5 分）】

組別(分)	人數(人)	
30 ~ 40	4	
40 ~ 50	6	
50 ~ 60	8	
60 ~ 70	14	
70 ~ 80	10	
80 ~ 90	6	
90 ~ 100	2	

平均絕對離差

1.平均絕對離差(mean absolute deviation)：簡寫成 *MAD*

一組資料 x_1、x_2、…、x_n，算數平均數為 $\overline{X}$ ，則定義：$MAD = \sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$ 。

討論：$\sum_{i=1}^{n} \frac{x_i - \overline{X}}{n} =$ _____ 。

變異數與標準差

1. 變異數與標準差：

設 N 個母體 $x_1, x_2, x_3, \ldots, x_N$ ，母體平均 μ ，抽出 n 個樣本 $x_1, x_2, x_3, \ldots, x_n$ ，樣本平均數 $\overline{X}$ ，則

(1)未分組求變異數與標準差：

定義：母體變異數：$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - N\mu^2}{N}$

樣本變異數：$S^2=\dfrac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}{n-1}=\dfrac{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}{n-1}$

母體標準差：$\sigma=\sqrt{\dfrac{\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{N}}=\sqrt{\dfrac{\sum_{i=1}^{N}x_i^2-N\mu^2}{N}}$

樣本標準差：$S=\sqrt{\dfrac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}{n-1}}=\sqrt{\dfrac{\sum_{i=1}^{n}x_i^2-n\overline{X}^2}{n-1}}$

(2)已分組求變異數與標準差：

已組中點當該組資料代表，定義跟已分組資料相同。

變量 X（組中點）	x_1	x_2	.....	x_n	總計
f	f_1	f_2		f_n	n

2. 資料的平移及伸縮對變異數與標準差的影響：

設一群數值資料 x_1、x_2、x_3、...、x_n，標準差為 S，設 A,B 為異於 0 的實數，則：

(1) x_i+A 的變異數為_____標準差為_____。

(2) Bx_i 的變異數為_____標準差為_____。

(3) Bx_i+A 的變異數為_____標準差為_____。

證明：

3.標準差的簡化計算：

為了簡化計算，通常我們會將資料 $x_1,x_2,x_3,\ldots,x_n$ 轉換成資料 $y_1,y_2,y_3,\ldots,y_n$，而 $y_i=\dfrac{1}{h}\left(x_i-A\right)$，$A$ 常為某組的組中點，這樣一來，Y 的數值會變得比較容易計算，先算出 $y_1,y_2,y_3,\ldots,y_n$ 的標準差 S_Y，就可計算 $x_1,x_2,x_3,\ldots,x_n$ 的標準差 $S_x=|h|S_y$。

・討論：$S_x = 0$的充要條件為何？

例 15：求母體資料 10 筆 86 ,95 , 73 , 80 , 67 , 79 , 61 , 85 , 62 , 82 的標準差。 【$\sqrt{110.4}$】

例 16：求例 5 中的資料抽 5 筆資料 73 , 80 , 67 , 79 , 61 的標準差。

例 17：求下列某班學生數學成績標準差。 【$\sqrt{239.36}$】

例 18：上例題的數學成績，抽 10 筆資料，其次數分配表如下，求數學成績標準差。

組別(分)	人數(人) fi	xi				
30 ~ 40	0					
40 ~ 50	2					
50 ~ 60	2					
60 ~ 70	3					
70 ~ 80	2					
80 ~ 90	1					
90 ~ 100	0					
total						

例 19：由臺北市選民抽出 $n=1050$ 位，訪問他們是否要投給候選人甲，結果有 420 位選民要投給候選人甲，設第 i 位受訪者資料為 x_i，即若第 i 位投給候選人甲，則 $x_i=1$，否則 $x_i=0$，求候選人甲樣本得票率與樣本標準差。 【$\sqrt{0.2402}$】

例 20：5 位學生參加歌唱比賽，甲、乙、丙、丁四位評審評分(1～10 分）如下：

(1) 求甲、乙、丙、丁四位評審的平均評分與標準差。 【0,0,$\sqrt{2}$,2$\sqrt{2}$】

(2) 甲、乙、丙、丁四位評審何者評分最有影響力？

【因評審丁的評分標準差最大，所以評審丁最有影響力】

		參賽者				
		1	2	3	4	5
評審	甲	10	10	10	10	10
	乙	1	1	1	1	1
	丙	1	2	3	4	5
	丁	1	3	9	7	5

例 21：某校模擬考試，第二類組 200 位學生的數學平均成績為 71 分，標準差 4 分，第三 類組 100 位學生的數學平均成績為 77 分，標準差為 5 分，求此 300 人的數學平均成績及標準差。

【73，5.2 (3$\sqrt{3}$)】

例 22：某次考試，平均為 30 分，全距 35，四分位差 15，標準差為 6，因學生很認真，老師將每個人的成績乘以 2 倍後，再加 10 分，求調分後成績的

（1）平均（2）全距（3）四分位差 （4）變異數（5）標準差。

【（1）70 （2）70 （3）30 （4）144 （5） 12】

4-2 二維數據分析

散布圖

1. 相關的意義：

兩種或兩種以上變數間的相關係，稱之為相關。

初步觀察相關程度最常使用的方式是畫圖。

EX：身高與體重、生產量與價格、油價與物價間的關係。

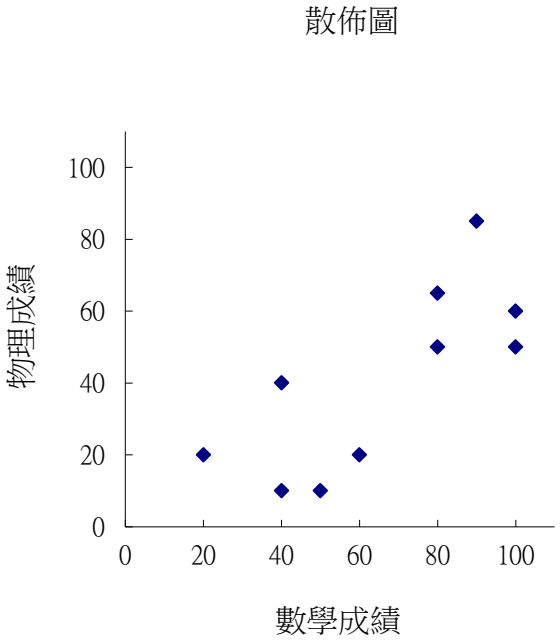
2. 散布圖：

設蒐集兩個變數 X、Y 的 n 筆資料(x_i, y_i)，$i=1,...,n$，將每筆(x_i, y_i)點畫在 XY 坐標平面上所得圖形稱為 Y 對 X 的**散布圖**．

・X 稱為**自變數**，Y 稱為**應變數**。

例 1：以下有 10 位學生的數學成績與物理成績資料，試畫散布圖。

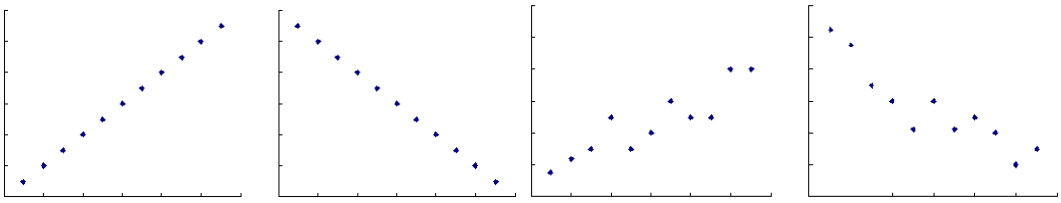
學生	數學成績 X	物理成績 Y
1	20	20
2	40	10
3	40	40
4	50	10
5	60	20
6	80	50
7	80	65
8	100	60
9	90	85
10	100	50



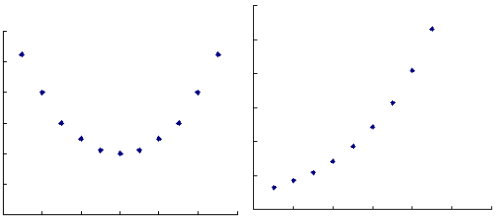
3. 散布圖的型態

(1)依相關的形態分：

・直線相關：兩個或兩個以上變數間的相互關係可用直線方程式適當的表示者。

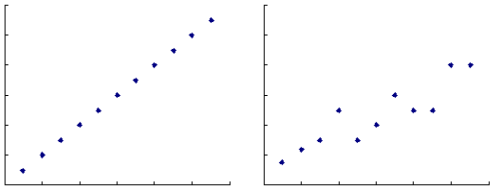


・曲線相關：兩個或兩個以上變數間的相互關係須用曲線方程式適當的表示者。

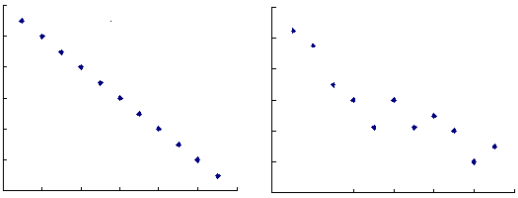


2. 依相關的方向分：

(1)正相關：變數間相對應的數值同增或同減，其變動方向一致者。



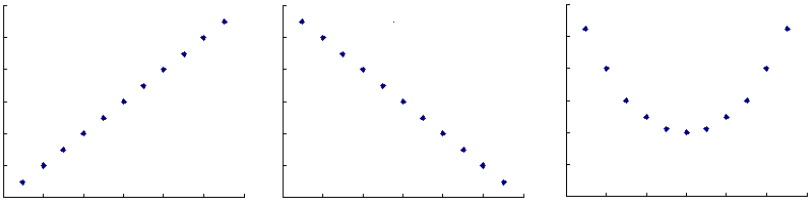
(2)負相關：變數間相對應的數值有此增彼減或此減彼增，其變動方向相反者。



3. 依相關的程度分：

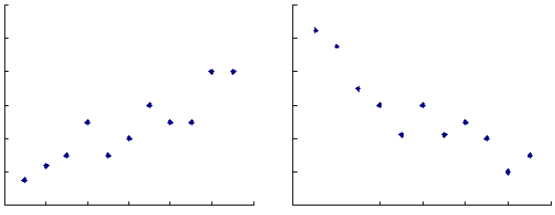
(1)完全相關：凡變數間的相關可以適當的用一條直線或一條曲線表示者。

【水平線、鉛直線例外】

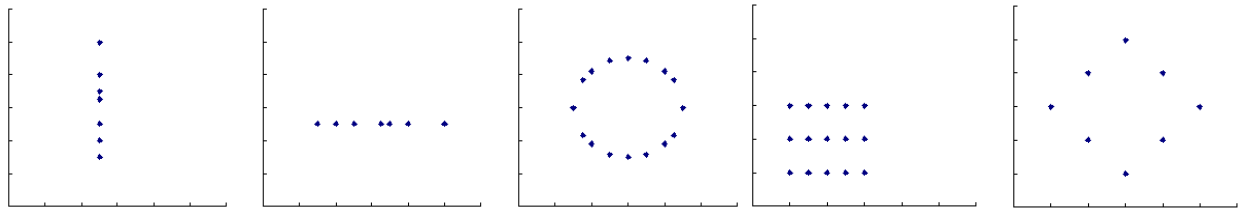


(2)不完全相關：變數間的相關雖不能以一條直線或一條曲線適當的表示，

但仍可以大致的表示者。



(3)零相關：變數間沒有關係。



相關係數

1. **相關係數**的意義：

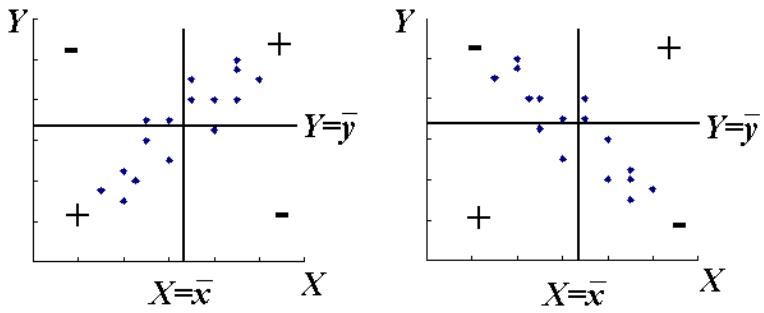
相關係數是用以測量兩變數間的**直線相關程度大小**及正負方向的量數，

以 r 表示之。

其計算公式如下：

有 n 對資料$(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)$，則

$$\text{相關係數 } r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$



例 2：某肥皂廠商欲推出一種新產品，在上市前以不同的單價 x(單位：十元)調查市場的需求量 y (單位：萬盒)。調查結果如下，問 x 和 y 的相關係數最接近下列那一個值？[84 學測]【-0.8】

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
8	11					
9	12					
10	10					
11	8					
12	9					

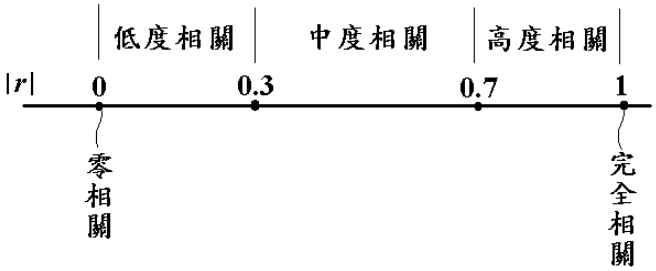
2. 性質：

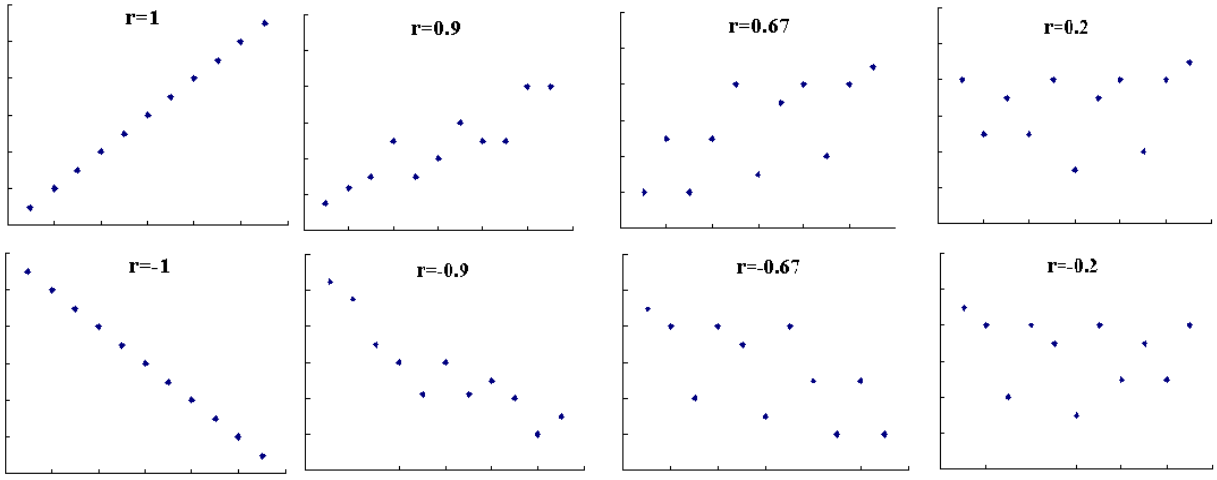
(1) $-1 \le r \le 1$。

證：

(2) 相關係數的性質：$|r|$愈大表示兩變量之間的相關程度愈大。

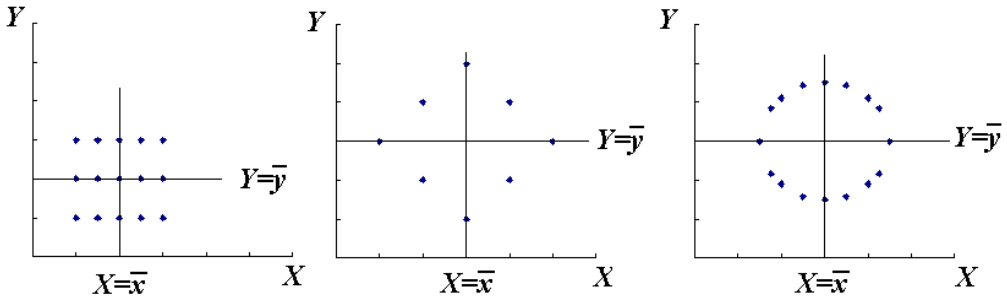
1. $|r| <= 1$ 表完全相關【$r = 1$ 表完全正相關; $r = -1$ 表完全負相關】。
2. $0.7 \le |r| < 1$ 表高度相關。
3. $0.3 \le |r| < 0.7$ 表中度相關 。
4. $0 < |r| < 0.3$ 表低度相關。





【完全相關】 【高度相關】 【中度相關】 【低度相關】

5 r =0 表零相關，可能情況：



(3) 相關係數的另一種表示：

$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$ ， $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ ， $S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})$ ，

$$\text{相關係數 } r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\cdot\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n(\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}y_i^2-n(\bar{y})^2}}$$

(4) 設 r_{xy} 表 n 對資料$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$的相關係數，

若 $u_i=ax_i+b$ ， $v_i=cy_i+d$ ，i=1,…,n，則(u_i,v_i)的相關係數

$$r_{uv}=\frac{ac}{|ac|}r_{xy}$$

證：

例 3：有學生十人(甲、乙、……、癸)，其期考數學成績與該學期數學課缺課數如下表所示：設兩者的相關係數為 r，則

(A) $-1\leqq r\leqq -0.6$，(B) $-0.6<r<-0.2$，(C) $-0.2\leqq r\leqq 0.2$，(D) $0.2<r<0.6$，(E) $0.6\leqq r\leqq 1$。[86 自]

【A】

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	100					
2	90					
3	90					
3	80					
4	70					
3	70					
5	60					

6	60					
3	80					
0	100					

例4：高三某次數學與物理競試後，抽出20位同學，他們的成績 (x_i，y_i)，i=1，2，…，20，(x_i，y_i分別表示第i個同學數學與物理成績)，整理得下面的數值：

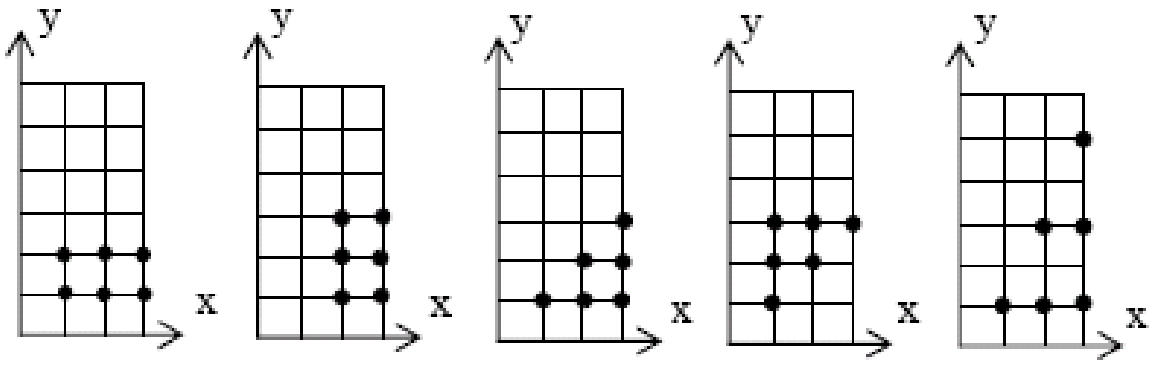
$\sum_{i=1}^{20} x_i$　160，$\sum_{i=1}^{20} y_i$　1160，$\sum_{i=1}^{20} x_i^2$　1302，$\sum_{i=1}^{20} y_i^2$　1320，$\sum_{i=1}^{20} x_i y_i$　，

(1) 計算這20位同學數學與物理競試成績的相關係數。　【$\frac{5\sqrt{55}}{44} \doteq 0.84$】

(2) 若每個人的各科成績都減1分，則相關係數為何？　【不變】

例 5：下圖中，有五組數據，每組各有 A、B、C、D、E、F 等六個資料點。設各組的相關係數分別為 r_1，r_2，r_3，r_4，r_5，則下列關係式何者為真？

(A) $r_1 = r_2$，(B) $r_2 < r_3$，(C) $r_3 > r_4$，(D) $r_3 < r_5$，(E) $r_4 = r_5$ [86 學測]　【ABE 】



例 6：令 X 代表每個高中生平均每天研讀數學的時間(以小時計)，則 W=7(24-X)代表每個高中生平均

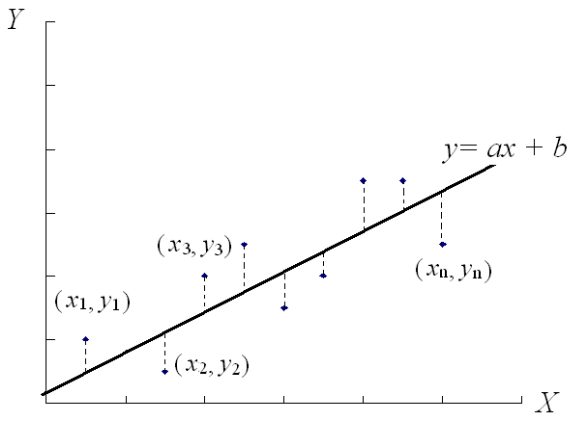
每週花在研讀數學以外的時間。令 Y 代表每個高中生數學學科能力測驗的成績。設 X，Y 之相關係數為 R_{XY}，W，Y 之相關係數為 R_{WY}，則 R_{XY} 與 R_{WY} 兩數之間的關係，下列選項何者為真？
(1) $R_{WY}=7(24-R_{XY})$，(2) $R_{WY}=7R_{XY}$，(3) $R_{WY}=-7R_{XY}$，(4) $R_{WY}=R_{XY}$，(5) $R_{WY}=-R_{XY}$。[90 學測]　【5】

最適直線方程式

1. 最小平方法：

對一組資料(x_i, y_i)，當由變數 Y(應變數)對變數 X(自變數)的散布圖看出此兩變數間有線性的圖案時，如何找出此條代表的直線呢?

假設 $y=a+bx$ 為代表的直線，



s1. 先求出殘差 $e_i=(a+bx_i-y_i)$

s2. 再求殘差平方和 $e_i^2=\sum_{i=1}^{n}(a+bx_i-y_i)^2$

s3. 當殘差平方和愈小，表示以此直線擬合此組資料愈好這種想法就稱為**最小平方法**

所謂最小平方法就是尋求使 $e_i^2=\sum_{i=1}^{n}(a+bx_i-y_i)^2$ 最小時的 a 、 b

2. 最適直線方程式(迴歸式)：

設有 n 對資料(x_i, y_i)， $y=a+bx$ 為代表的直線

則當 $b=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$，$a=\bar{y}-b\bar{x}$時，殘差 e_i^2 的值為最小。

此直線 $y=a+bx$，稱為 Y 對 X 的最適直線方程式(Y 對 X 的迴歸式)。

證：

【重要】 Y 對 X 的最適直線方程式(Y 對 X 的迴歸式)兩個重要形式：

(1) $y-\bar{y}=\dfrac{S_{xy}}{S_{xx}}(x-\bar{x})$，其中 $S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$ ， $S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})$ 。

(2) $y-\bar{y}=r\dfrac{S_y}{S_x}(x-\bar{x})$，其中 S_x 表 x 的標準差，S_y 表 y 的標準差，

r 表相關係數。

3.迴歸預測：最適直線方程式預測別的情況。

假設已知迴歸方程式 $y=a+bx$ ，則給定 x_h 時的預測值 $y_h=a+bx_h$ 。

例 7：設有一組資料：試求 y 對 x 的迴歸式。 【0.84,y=0.8+3.4x】

X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	4					
3	11					
5	14					
4	20					
2	6					

例 8：1. 求例 4 中出 Y 對 X 之最適合直線方程式。 【y=1.13x+1.04】

2. 利用此直線方程式預測，若學生數學考 60 分則物理考幾分？ 【66.76】

例 9：若一組資料 x,y 的相關係數為 0.6，x 的平均數為 60，x 的標準差為 4，y 的平均數為 42， y 的標準差為 3，則：

(1) 求 y 對 x 的迴歸式。【y=15+0.45x】 (2) 求 x 對 y 的迴歸式。【x=26.4+0.8y】