4-2　二維數據分析

我們除了分析單一數據資料外，常常也需要一起考慮兩組數據，去發現它們之間可能存在的因果關係，以及對彼此造成的影響，這就是二維數據分析。本章將討論「散佈圖」、「相關係數」與「迴歸直線」等工具，探討、解釋並預測兩組數據間的相互關係。

散佈圖

具有兩個變量的數據，稱為**二維數據**，這在現代生活與科學中處處可見。

例如：(身體，體重)，(容量，體積)，(風力，雨量)，(數學成績，國文成績)，(BMI 指數，血壓）等等。

通常我們將兩個可能相關的變量數據以點標示在坐標平面上，第一個變量當作 x 坐標，第二個變量當作 y 坐標，選取適當刻度後，將每一組資料（x_i，y_i）描繪在坐標平面上，這樣所得的圖形稱為**散佈圖**。

例題 1----------

測量 7 位同學的身高與體重，結果如下表：

身高（公分）	172	160	162	164	170	168	166
體重（公斤）	60	50	52	58	62	56	54

試繪出其散佈圖。

隨堂練習----------

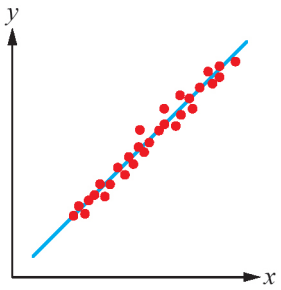
蒐集 7 筆有關年齡與血壓的二維數據，結果如下表，試繪出其散佈圖。

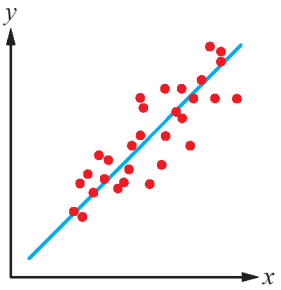
年齡（歲）	35	40	45	50	55	60	65
收縮壓（mmHg）	116	120	124	128	132	136	140

正相關、負相關、零相關

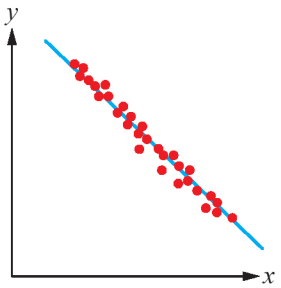
由散佈圖可以快速觀察出兩個變量之間是否有關係。

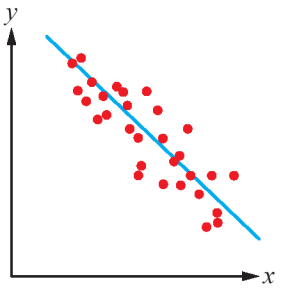
(1) **正相關**：兩個變量有一致的趨勢（同時增加或減少）。



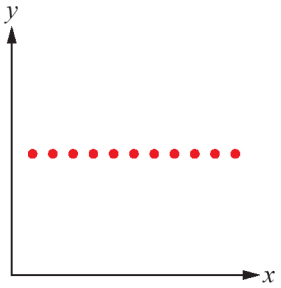


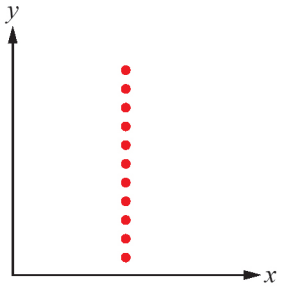
(2) **負相關**：兩個變量趨勢相反，一個增加（減少）則另一個就減少（增加）。

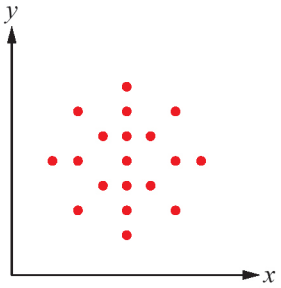




(3) **零相關**：一個變量的變化對另一個變量沒有影響。

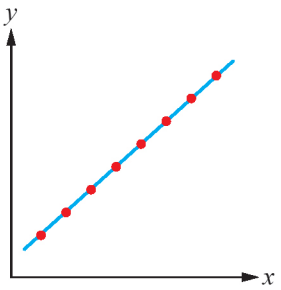


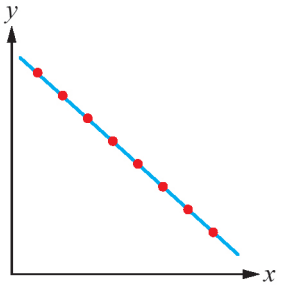




(4) **完全正相關**：資料全部在一條斜率為正的直線上。

(5) **完全負相關**：資料全部在一條斜率為負的直線上。





數據標準化

描繪散佈圖時，由於單位與刻度可以任意選定，同一組資料（x_i，y_i）的描繪結果可能差異很大。因此通常我們會先將數據標準化後再描繪出散佈圖。

※二維數據的標準化

設資料的第一個變量的平均數為 μ_x，標準差為 σ_x；第二個變量的平均數為 μ_y，標準差為 σ_y。則（x_i，y_i）的標準化數據為（u_i，v_i），其中

$$u_i=\frac{x_i-\mu_x}{\sigma_x},\ v_i=\frac{y_i-\mu_y}{\sigma_y}。$$

例題 2----------

承例題 1 的數據，試繪出標準化數據的散佈圖。

身高（公分）	172	160	162	164	170	168	166
體重（公斤）	60	50	52	58	62	56	54

解　由例題 1 的數據，經計算可得 $\mu_x=166$，$\sigma_x=4$，$\mu_y=56$，$\sigma_y=4$。標準化數據如下：

$u_i=\frac{x_i-\mu_x}{\sigma_x}$							
$v_i=\frac{y_i-\mu_y}{\sigma_y}$							

隨堂練習----------

例題 1 的隨堂練習中，

年齡（歲）	35	40	45	50	55	60	65

收縮壓（mmHg）	116	120	124	128	132	136	140

若年齡 $\mu_x=50$，$\sigma_x=10$，收縮壓 $\mu_y=128$，$\sigma_y=8$。試繪出標準化數據的散佈圖。

標準化數據有幾個特性：

1. 標準化數據的變量都沒有單位。
2. 標準化數據兩變量的平均數皆為 0，標準差皆為 1。

原來兩變量的平均值經過標準化後變成原點，兩軸的單位長就是兩變量的標準差。

例題 3----------

蒐集 7 筆年齡與睡眠時間的數據，結果如下表：

年齡（歲）	24	28	32	36	40	44	48
睡眠時間（時）	8	7.8	7.6	7.4	7.2	7	6.8

試繪出標準化數據的散佈圖。

解　以 x 坐標表示年齡、y 坐標表示睡眠時間。則經過標準化的數據如下：

$u_i=\frac{x_i-\mu_x}{\sigma_x}$							
$v_i=\frac{y_i-\mu_y}{\sigma_y}$							

隨堂練習----------

如例題 3，試選取適當刻度繪出原始資料的散佈圖。

標準化數據的結論：

1. 若原始數據落在斜率為正的直線上，則標準化後的數據必落在通過原點、斜率為 1 的直線上。
2. 若原始數據落在斜率為負的直線上，則標準化後的數據必落在通過原點、斜率為 -1 的直線上。

例題 4----------

設二維數據原始資料為（x_i，y_i），標準化的數據為（u_i，v_i）。試證明：

若原始數據落在直線 $y=3x+2$ 上，則標準化後的數據落在直線 $v=u$ 上。

隨堂練習----------

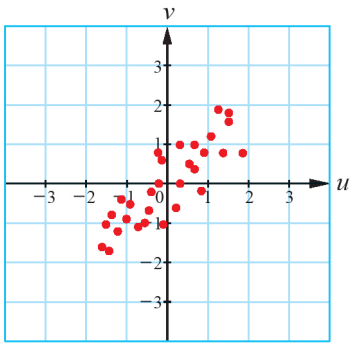
試證明：若原始數據落在直線 $y=-5x+1$ 上，則標準化後的數據落在直線 $v=-u$ 上。

相關係數 （ 衡量兩變量相關的程度 ）

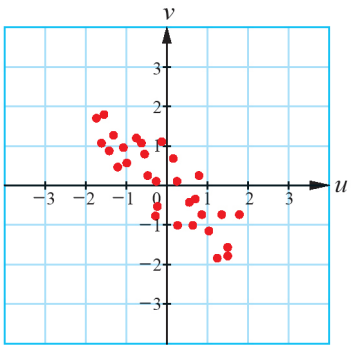
將兩組數據中的其中一組進行平移或伸縮，並不影響兩變量之間的相對關係。所以為了有效描述兩變量之間的關係，我們需將兩組數據標準化，亦即使各組數據的平均值為 0，標準差為 1。將數據標準化之後，在散佈圖上兩變量的平均值會成為原點。設每一筆資料的標準化數據為（u_i，v_i）。當點（u_i，v_i）在第一、三象限時，有 $u_iv_i>0$；而當點（u_i，v_i）在第二、四象限時，有 $u_iv_i<0$。

所以，如果 $\sum_{i=1}^{n} u_iv_i$ 這個值是正的，表示落在第一、三象限的點多，圖形會是右上左下的趨勢（正相關），如圖 5，而且正愈多，表示趨勢愈強。同理，如果 $\sum_{i=1}^{n} u_iv_i$ 這個值是負的，

那表示落在第二、四象限的點多，圖形會是左上右下的趨勢（負相關），如圖 6，而且負愈多，表示趨勢愈強。



正相關



負相關

因此，$\sum_{i=1}^{n} u_i v_i$ 可以用來衡量相關程度。為了消弭資料個數的影響，我們除掉資料個數，得到 $\frac{1}{n}\left(\sum_{i=1}^{n} u_i v_i\right)$。此即**相關係數**的定義，通常記為 r。

※相關係數

標準化數據（u_i，v_i），$i=1$，2，$\cdots$，n **的相關係數定義為**

$$r=\frac{1}{n}\left(\sum_{i=1}^{n} u_i v_i\right)。$$

例題 5----------

例題 2 的標準化數據如下表：

$u_i=\frac{x_i-\mu_x}{\sigma_x}$	1.5	−1.5	−1	−0.5	1	0.5	0
$v_i=\frac{y_i-\mu_y}{\sigma_y}$	1	−1.5	−1	0.5	1.5	0	−0.5

試求其相關係數。（取到小數點後第四位）

解　直接計算得

$$r=\frac{1}{7}\left(\sum_{i=1}^{n} u_i v_i\right)$$

$$=\frac{1}{7}(1.5+2.25+1-0.25+1.5+0+0)$$

$$=\frac{6}{7}\approx 0.8571,$$

即相關係數約為 0.8571。

隨堂練習----------

試求以下標準化數據的相關係數。（取到小數點後第四位）

$u_i=\frac{x_i-\mu_x}{\sigma_x}$	-2	-0.6	-0.3	0.3	0.5	1	1.1
$v_i=\frac{y_i-\mu_y}{\sigma_y}$	2	$-\frac{1}{6}$	$\frac{4}{6}$	$-\frac{7}{6}$	$\frac{1}{6}$	$-\frac{5}{6}$	$-\frac{4}{6}$

※相關係數

原始數據資料(x_i, y_i)，$i=1, 2, \cdots, n$ 的相關係數定義為

$$r=\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{(x_i-\mu_x)^2\cdot(y_i-\mu_y)^2}}。$$

這個式子的好處是可以省去標準化的步驟。

如前述理由，在未標準化的散佈圖上，以(μ_x, μ_y)為原點將坐標平面分成四個象限，

若 $r>0$ 時，表示在第一、三象限的點較多（正相關）；

若 $r<0$ 時，表示在第二、四象限的點較多（負相關）。

利用高二會學到的柯西不等式，我們可以證明$-1\le r\le 1$（可參考附錄），且其性質如下：

(1) $r>0$ 表兩變量正相關，$r<0$ 表負相關，$r=0$ 表零相關。

(2) $r=1$ 表兩變量完全正相關，$r=-1$ 表完全負相關。

(3) $|r|$越大表示兩變量間的相關程度愈強。

相關係數很高，這兩個變量也不一定有因果關係，需要對整體狀況有進一步的了解之後，才能下定論。

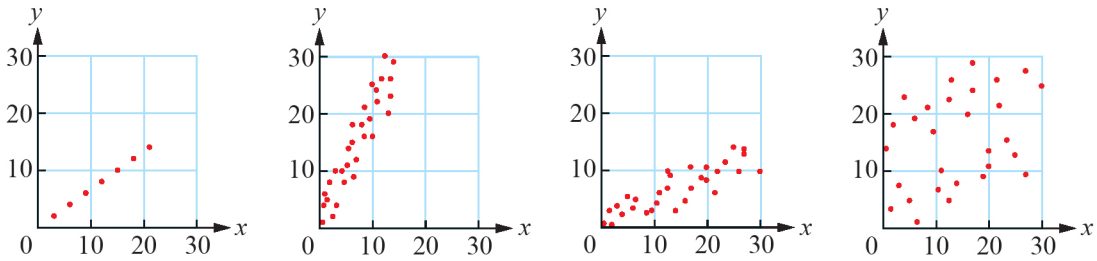
因為 $u_i=\frac{x_i-\mu_x}{\sigma_x}$，$v_i=\frac{y_i-\mu_y}{\sigma_y}$，故相關係數亦可化為用原始數據來計算：

$$r=\frac{1}{n}\left(\sum_{i=1}^{n}u_iv_i\right)=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)$$

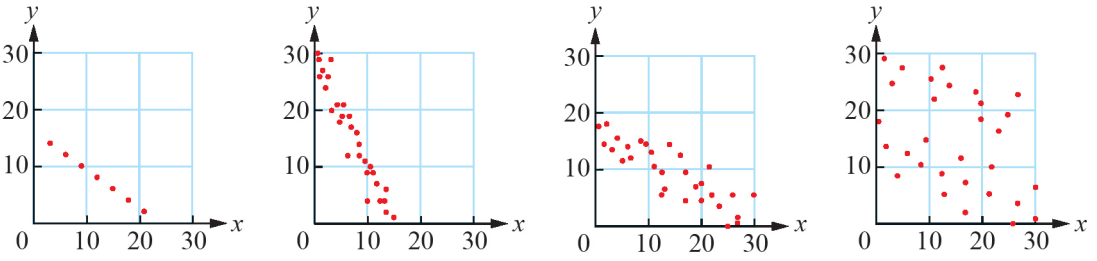
$$=\frac{1}{n}\cdot\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\cdot\sigma_y}$$

$$=\frac{1}{n}\cdot\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\frac{1}{n}(x_i-\mu_x)^2}\cdot\sqrt{\frac{1}{n}(y_i-\mu_y)^2}}$$

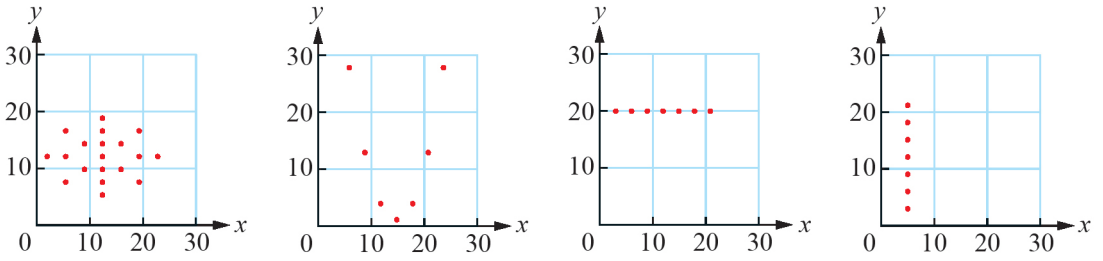
$$= \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{(x_i - \mu_x)^2 \cdot (y_i - \mu_y)^2}}$$



$r = 1$ 完全正相關　　$r = 0.9$ 高度正相關　　$r = 0.7$ 高度正相關　　$r = 0.3$ 低度正相關



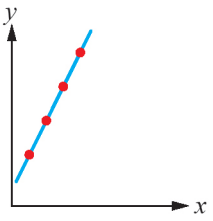
$r = -1$ 完全負相關　　$r = -0.9$ 高度負相關　　$r = -0.7$ 高度負相關　　$r = -0.3$ 低度負相關

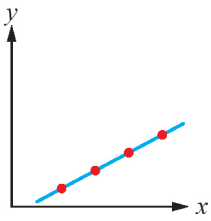


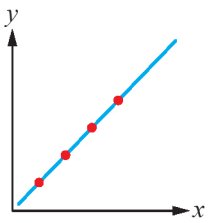
$r = 0$ 零相關　　$r = 0$ 零相關　　$r = 0$ 零相關　　$r = 0$ 零相關

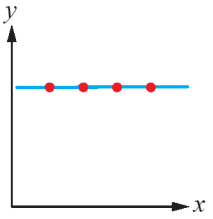
隨堂練習----------

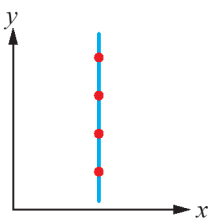
以下 5 組原始數據的散佈圖，試問哪些相關係數為 1？

(A) 

(B) 

(C) 

(D) 

(E) 

最小平方法與迴歸直線

本節我們要找一條最適合代表兩變量之間關係的直線 L，稱為**迴歸直線**（或**最適直線**），以當作預測的依據。尋找直線 L 的方法稱為**最小平方法**，是由數學家高斯所提出的。關鍵的想法是：

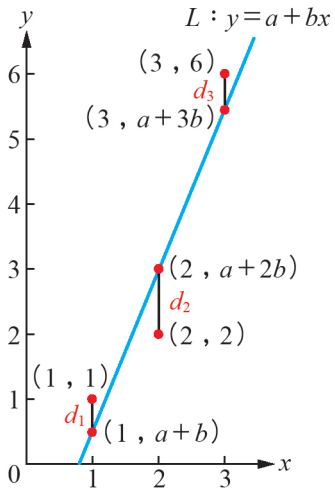
使得資料點到直線 L 的鉛垂距離的平方和最小（因此稱為最小平方法）。

例題 6----------

散佈圖上有資料（1，1），（2，2），（3，6），試用最小平方法求迴歸直線方程式。

解

（　　）



隨堂練習----------

給定 3 個二維數據分別是（1，0），（2，2），（3，7），試利用最小平方法求迴歸直線方程式。

當數據很多時，仿上述方法求迴歸直線方程式的計算量相當大。但若先考慮標準化後的數據，設其迴歸直線方程式為 $Y=a+bX$，則可以證得美妙的結果：迴歸直線通過原點，且斜率恰好就是相關係數。

※標準化數據的迴歸直線方程式

設 $(u_i，v_i)$，$i=1，2，\cdots，n$，是標準化後的數據，則迴歸直線方程式為

$$Y=rX，\text{其中 } r \text{ 為相關係數。}$$

此式的優點是不需要計算相關係數與標準差。

※二維數據的迴歸直線方程式

設 $(x_i，y_i)$，$i=1，2，\cdots，n$ 為二維數據，則迴歸直線方程式為

$y-\mu_y=r\dfrac{\sigma_y}{\sigma_x}(x-\mu_x)$，其中 r 為相關係數。此迴歸直線方程式

亦可寫為 $y-\mu_y=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{(x_i-\mu_x)^2}\cdot(x-\mu_x)$。

例題 7----------

下表為五位同學的身高體重：

身高（公分）	160	164	168	172	176
體重（公斤）	50	52	48	56	54

令 x 表身高，y 表示體重，試求迴歸直線方程式。

解　計算得算術平均數 $\mu_x=168$，$\mu_y=52$。製作表格如下：

x	y	$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)(y-\mu_y)$	$(x-\mu_x)^2$	$(y-\mu_y)^2$

迴歸直線的意義，在於二維數據資料的分布，在某種測量的方式之下，最能代表兩組變量相互關係的線型方程式。我們利用已知的二維數據資料求得迴歸直線後，通常會使用此直線方程式作為模型，計算在 x 變量改變的情形下，y 變量可能會呈現如何的數據，以作為預測與決策的根據。

例題 8----------

下表為每公頃的土地上，使用肥料量（公斤）與產量（公斤）的關係：

肥料量（公斤）	280	300	320	340	360	380	400
產量（公斤）	7150	7100	7200	7250	7350	7400	7300

試問每公頃施肥量為 350 公斤時，產量約為多少？

解　假設 x 表示肥料量，y 表示產量，計算得算術平均數 $\mu_x=340$、$\mu_y=7250$，製作表格如下：

x	y	$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)(y-\mu_y)$	$(x-\mu_x)^2$	$(y-\mu_y)^2$

隨堂練習----------

某地區近 9 年每人每日垃圾清運量與資源回收率的資料如下，若每人每日垃圾清運量為

0.53 公斤時，試預測資源回收率為何？

每人每日垃圾清運量（公斤）	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
資源回收率（％）	15	17	20	22	25	27	29	33	37

習　題　4-2

一、基本題

1. 下列哪些選項的敘述是正確的？
   (A) 相關係數 r 一定滿足 $-1 \le r \le 1$
   (B) 若兩變數成直線關係，則相關係數為 1
   (C)（x，y）的相關係數 r_{xy} 與（y，x）的相關係數 r_{yx} 相同
   (D) 二維數據的單位改變之後，相關係數也會改變
   (E) 將數據標準化，不會改變相關係數

2. 已知一組二維數據如下表所示：

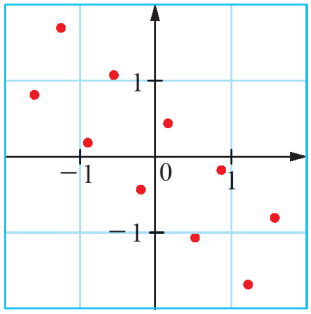
x	2	6	8	10	14
y	3	6	5	9	7

   (1) 試求其標準化數據。
   (2) 試繪出標準化數據散佈圖。
   (3) 求相關係數。

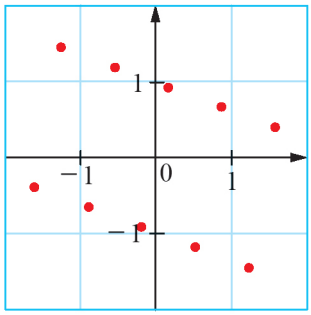
3. 給定 4 個二維數據分別是（1，2），（3，6），（5，4），（7，8），試利用最小平方法求最適直線方程式。

4. 設 A，B，C 是三組資料，其標準化散佈圖由左至右排列如下：

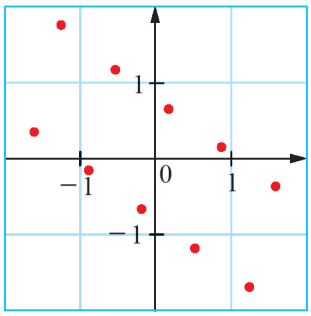
A 的標準化散佈圖



B 的標準化散佈圖



C 的標準化散佈圖



若 A 組資料與 B 組資料的相關係數分別為 -0.8 與 -0.2，則下列何者最可能是 C 組資料的相關係數？

(A) -1　(B) -0.9　(C) -0.6　(D) -0.1　(E) 0

5. 某一公司行銷部門蒐集廣告費（千元）與銷售量（千個）的資料如下表，假設 x 表示廣告費，y 表示銷售量。

廣告 (x)	1	3	5	7	9	11	13
銷售 (y)	1	3	2	5	4	7	6

試求：

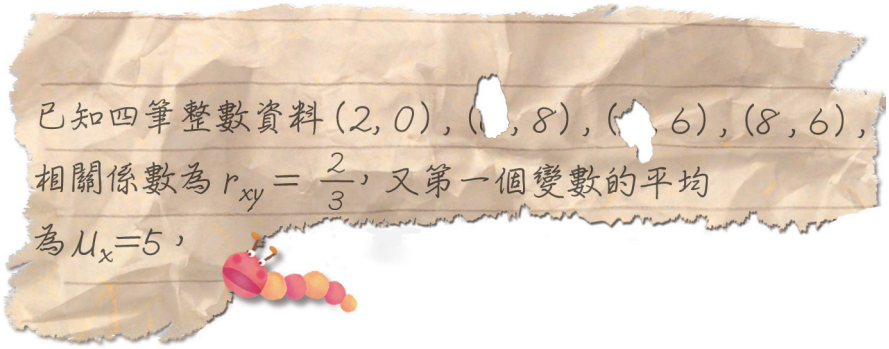
(1) y 對 x 的迴歸直線方程式。

(2) 利用迴歸直線方程式，預測當廣告費用是 10（千元）時，銷售量大約為多少（千個）？（取到小數點後第二位）

二、進階題

6. 已知三筆資料（5，3），（1，t），（3，1）y 對 x 的迴歸直線方程式是 $y=\frac{5}{4}+\frac{1}{4}x$，試求 t 值。

7. 小璿發現爺爺的古老數學筆記本，上面有這樣的題目：



小璿想了一下說，我知道被蟲蛀掉的數是什麼。試求出被蟲蛀掉的兩個數。

8. 令 x 表示國民每天平均睡覺的時間，$y=24-x$ 為國民每天平均醒著的時間，w 表示國民平均生產毛額。令 r_{xw} 為 x, w 的相關係數，令 r_{yw} 為 y, w 的相關係數。試將 r_{yw} 用 r_{xw} 表示。

9. 有一組二維數據（x_i，y_i），$i=1$，2，…，n，若已知 y 對 x 的迴歸直線方程式為 $y=2x+1$，且已知平均數 $\mu_x=2$，$\mu_y=5$，標準差 $\sigma_x=3$，$\sigma_y=7$。試求 x 與 y 的相關係數。

三、挑戰題

10. (1) 假設二維數據（x_i，y_i）之相關係數為 r。令 $x_i'=ax_i+b$ 與 $y_i'=cy_i+d$，試證明新的二維數據（x_i'，y_i'）之相關係數為 $\frac{ac}{|ac|}r$。

(2) 設二維數據（x_i，y_i）之相關係數為 $r=0.6$，則（$2x_i+3$，$5y_i+4$）之相關係數為多少？

(3) 設二維數據（x_i，y_i）之相關係數為 $r=0.6$，則（$2x_i+3$，$-5y_i+4$）之相關係數為多

少？

(4) 試說明為什麼將二維數據（x_i，y_i）的資料標準化後，相關係數不會改變。

第 4 章　綜合演練

1. 有 10 隻各式造型的 kitty 布偶，為增加其價值，每個布偶外加一個重量 200 克的包裝盒，則包裝前與包裝後，布偶重量的統計數值不會改變的有哪些？
(A) 算術平均數　(B) 中位數　(C) 眾數　(D) 變異數　(E) 標準差

2. 健身房有 10 位同學，每人左、右手各拿著一個等重的啞鈴。若每人右手所持的啞鈴重量的算術平均數是 8 磅，標準差是 2 磅。則所有啞鈴重量的算術平均數是多少？標準差是多少？

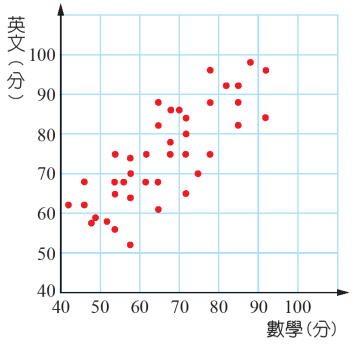
3. 小璿收到某次段考的成績單如下，試問小璿的班級排名較好是哪一科目？

科目 / 姓名	國文	英文	數學	歷史	地理
…	…	…	…	…	…
小璿	80	75	70	80	82
…	…	…	…	…	…
各科平均	76	80	50	80	88
標準差	4	5	10	5	4

4. 下表為 10 位同學參加學科能力測驗的數學科成績，其中 A，B 兩位同學的成績因印刷油污看不清楚。已知 10 位同學的算術平均數為 11 級分，變異數為 3 級分；又 A 的成績比 B 高，試問 A，B 兩位同學的成績分別為何？

姓名	A	B	C	D	E	F	G	H	I	J
成績			9	10	12	11	14	12	10	11

5. 高一甲班 40 人某次考試數學（橫軸）與英文（縱軸）成績之散佈圖如右，每個點代表一位學生的成績。若及格標準為 60 分，請問下列哪些選項是正確的？



(A) 兩科都不及格的學生有 5 位

(B) 數學的中位數大於英文的中位數
(C) 每位同學兩科成績總和都超過 100 分
(D) 數學的標準差大於英文的標準差
(E) 若以最小平方法決定數據集中趨勢的直線方程式，則該直線的斜率大於 0

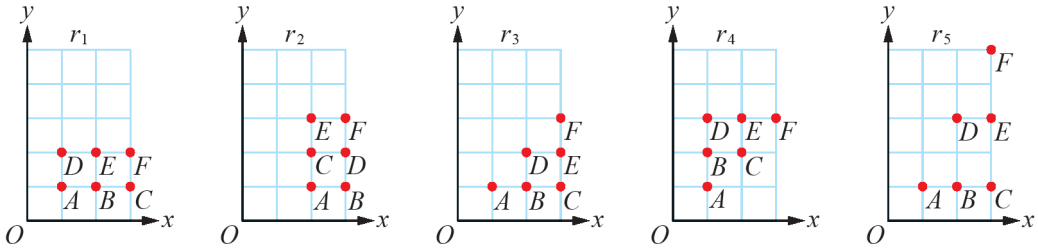
6. (1) 有一組二維數據（x_i，y_i），$i=1$，2，…，n，若已知平均數 $\mu_x=2$，$\mu_y=4$，標準差 $\sigma_x=3$，$\sigma_y=8$，且 x 與 y 的相關係數為 $-\frac{3}{4}$，試求 y 對 x 的迴歸直線方程式？

(2) 有一組二維數據（x_i，y_i），$i=1$，2，…，n，若已知平均數 $\mu_x=5$ $\mu_y=3$，標準差 $\sigma_x=4$，$\sigma_y=10$，且 y 與 x 的迴歸直線過點（3，7），試求 x 與 y 的相關係數？

7. 高一某班 40 位學生，數學科第一次段考、第二次段考分別以 x_i，y_i（其中 $i=1$，2，…，40）表示，且每位學生的成績用 0 至 100 評分。若這兩次段考數學科成績的相關係數為 0.6，試問下列哪些選項是正確的？

(A) $x-1$ 與 $y+2$ 的相關係數仍為 0.6
(B) $2x$ 與 $3y$ 的相關係數仍為 0.6
(C) 若 $x'=\frac{x-\mu_x}{\sigma_x}$、$y'=\frac{y-\mu_y}{\sigma_y}$，其中 μ_x，μ_y 分別為 x，y 的平均數，σ_x，σ_y 分別為 x，y 的標準差，用 x' 與 y' 的相關係數為 0.6
(D) $2x-1$ 與 $3y$ 的相關係數仍為 0.6
(E) $2x-1$ 與 $-3y+2$ 的相關係數仍為 0.6

8. 下圖中，有五組數據，每組各有 A，B，C，D，E，F 等六個資料點，設各組的相關係數由左而右分別為 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) $r_4=r_5$

9. 某班的 50 名學生參加一項考試，考題共有 100 題，全為是非題，計分方法共有 X，Y 兩種；若某學生有 R 題答對，W 題未答對（含答錯或未答），則 $X=R$、$Y=R-\frac{W}{5}$，試問下列敘述哪些是正確的？

(A) 同一學生的 Y 分數不可能大於 X 分數
(B) 全班 Y 分數的算術平均數不可能大於 X 分數的算術平均數
(C) 任兩學生 Y 分數的差之絕對值不可能大於 X 分數的差之絕對值

(D) 用 X 分數將全班排名次的結果與用 Y 分數排名次是完全相同的

(E) 兩種分數的相關係數為 1