

## §1-3 變異係數與相關係數

### (甲)變異係數

例子一：

設某次段考，高三某班的國文成績的算術平均數與標準差分別為 80 分、10 分；英文成績的算術平均數與標準差分別為 60 分、10 分；雖然國文與英文的標準差相等，如果我們得到結論是國文與英文成績的差異程度一樣，顯然不合理。現在我們比較兩科的標準差與算術平均數的比值：

國文科： $\frac{10}{80} = \frac{1}{8}$ ，英文科： $\frac{10}{60} = \frac{1}{6}$ ，從這兩科的比值來看，我們可以認為英文成績的差異會比國文成績的差異大。

例子二：

同時測量一張桌子的長度 10 次，10 次長度的算術平均數為 1.72 公尺，標準差為 0.04 公尺，若我們改變單位將公尺改為公分，算術平均數為 172 公分，標準差為 4 公分，若我們比較兩個標準差 0.04 公尺與 4 公分，雖然  $0.04 < 4$ ，但是若我們得到這兩筆資料的差異程度不同，這就會鬧笑話了！但是我們比較這兩筆資料標準差與算術平均數的比值： $\frac{0.04}{1.72} = \frac{4}{172}$ ，這就可以呈現出這兩筆資料的差異程度相同。

比較兩組或兩組以上的資料之差異時，通常採用一種相對的測度值作為比較的標準。因此無論兩筆資料的單位與取值範圍是否相同，若用算術平均數為基準，以標準差相對於算術平均數的比值來比較，就可以比較離散程度，比值愈大表示資料間的差異也愈大。

(1)變異係數(CV)的定義：

變異係數的定義： $CV = \frac{S_x}{\bar{x}} \times 100\%$ ， $S_x$ 為標準差， $\bar{x}$ 代表算術平均數。

CV的意義是計算標準差相對於算術平均數的百分比。

百分比越大，代表資料越分散。

[例題1] 某校高三有兩班，甲班學生身高平均值為 168.5 公分，標準差為 7.2 公分；乙班學生身高平均值為 159.6 公分，標準差為 4.8 公分。試問那一班學生身高較懸殊？

[解法]：

$$(CV)_{\text{甲}} = \frac{7.2}{168.5} \times 100\% = 4.27\%$$

$$(CV)_{\text{乙}} = \frac{4.8}{159.6} \times 100\% = 3.01\%$$

$\Rightarrow (CV)_{\text{甲}} > (CV)_{\text{乙}} \Rightarrow$  甲班學生身高差異較乙班大。

(練習1) 第一次月考，甲、乙兩班數學的平均成績分別為 80 分、70 分，標準差分別為 9 分，8 分，試問那一班的同學之程度比較平均？

Ans：甲班

(練習2) 某次數學考試，統計結果如下表：試求：

(1)全部 100 人的平均成績  $\bar{X}$  為多少分？

(2)那一班的程度比較平均？用數據說明。

(3)全部 100 人的標準差  $S$  為多少分？（至小數第三位再四捨五入）

Ans：(1)67 分 (2)乙班 (3)9.19 分

	甲班	乙班
平均成績	70 分	65 分
標準差	10 分	8 分
人數	40 人	60 人

## (乙) 相關係數

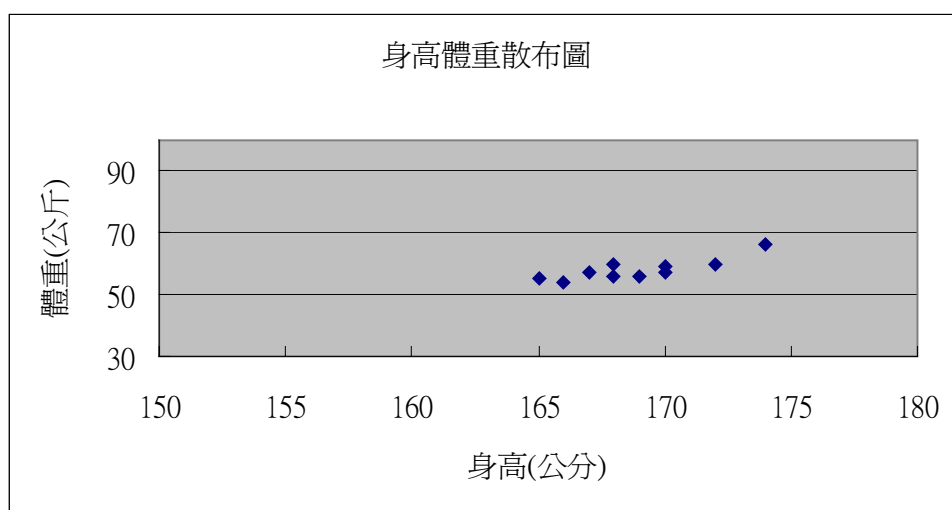
通常身高很高的人，體重不會太輕，物理成績高的學生，數學成績通常也不會很低，不管是身高、體重或是物理、數學成績，我們如何來衡量這兩個資料間的關係呢？可否由身高來預測體重，或是由數學成績來預測物理成績呢？

(1)散佈圖(scatter plot)：

設高三某班 10 位同學身高與體重成績的資料如下表所示：

學生編號	1	2	3	4	5	6	7	8	9	10
身高 X(公分)	168	172	170	166	174	167	169	165	170	168
體重 Y(公斤)	56	60	57	54	66	57	56	55	59	60

將兩個變數的數值資料數對畫在坐標平面上，以表明它們的分布情形的圖形，稱為**散佈圖**，散佈圖上的點稱為**樣本點**。

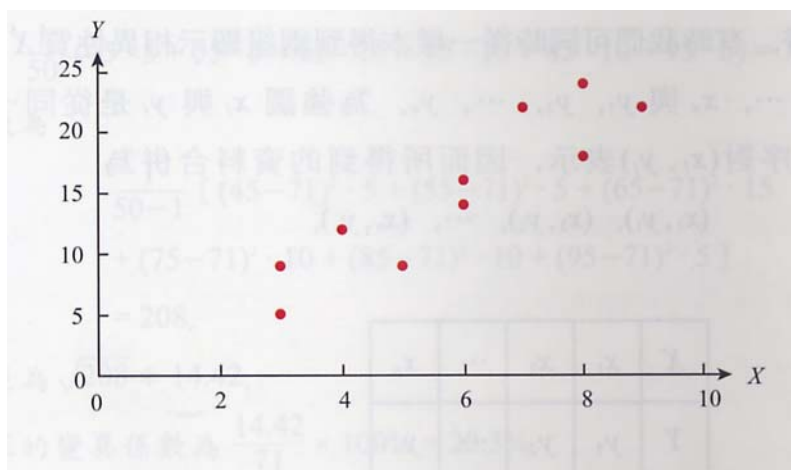


(2) 散布圖與相關程度：

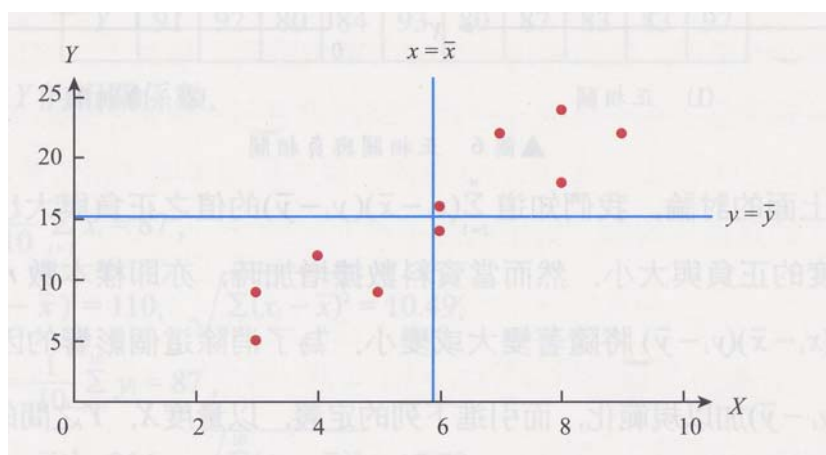
某種新藥的用量  $X$ (毫克數)與藥效期間  $Y$ (天數)的關係，  
經調查後得到資料如下表：

$X$	3	3	4	5	6	6	7	8	8	9
$Y$	9	5	12	9	14	16	22	18	24	22

畫出此資料的散布圖：



計算用藥的平均值  $\bar{x}=5.9$  毫克，藥效期間平均值  $\bar{y}=15.1$  天，若在散布圖中加畫  $x=5.9$ ， $y=15.1$  兩直線，則可將全圖分成四個區域：



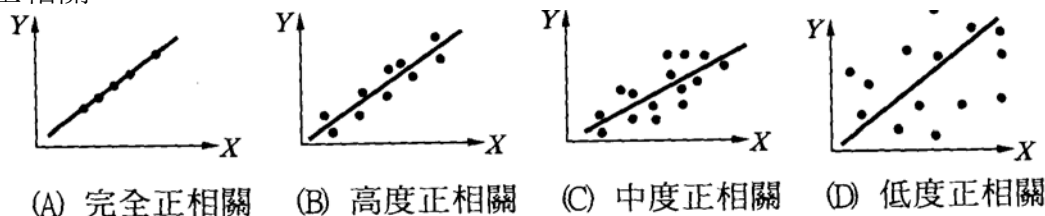
由上圖可以看出，除了(6,14)一點外，其餘的點都在右上區或左下區，這表示絕大多數的情形，若用藥超過平均值，則藥效期間亦超過平均值，反之亦然，換句話說，用藥量與藥效期間同時為增或同時為減，兩者之間是有某種程度的相關性。

一般而言，如果在散布圖中以  $y=\bar{y}$  為新的橫軸， $x=\bar{x}$  為新的縱軸，則可將全圖分成四個象限，在第一三象限內的點  $(x_i, y_i)$ ， $(x_i - \bar{x})(y_i - \bar{y})$  的值為正，；在第二四象限內的點  $(x_i, y_i)$ ， $(x_i - \bar{x})(y_i - \bar{y})$  的值為負，若資料內的樣本點  $(x_1, y_1)$ 、 $(x_2, y_2)$ 、...、 $(x_n, y_n)$  中，計算  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  的值：

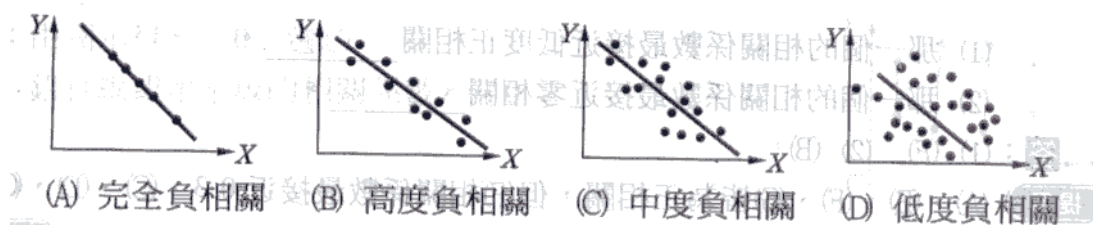
(a) 若  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) > 0$ ，則表示  $X$  與  $Y$  的變動趨勢大致相同，即同時為增或同時為減，我們稱兩者為**正相關**。

(b) 若  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) < 0$ ，則表示X與Y的變動趨勢大致相反，即此增彼減或此減彼增，我們稱兩者為**負相關**。

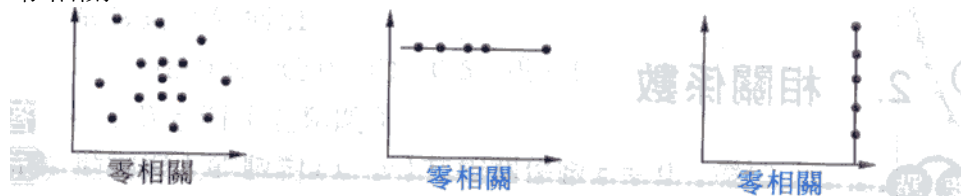
(c) 根據散布圖，有時可以大概判斷出兩個變數 X、Y 之間的相關情形：  
正相關：



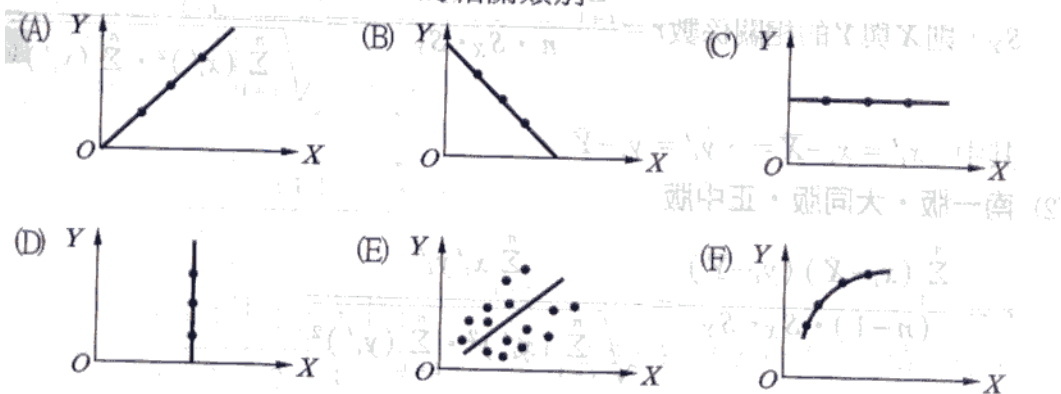
負相關：



零相關：



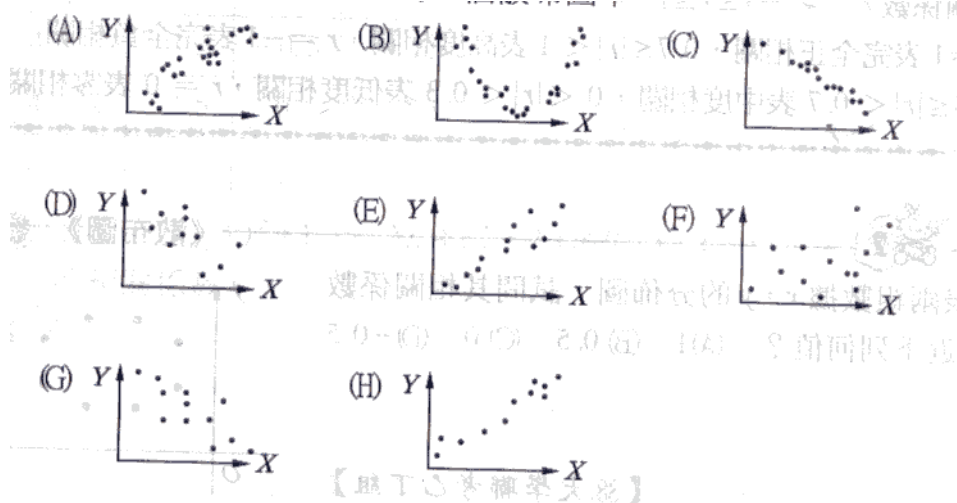
[例題2] 就下列圖形說明變量 X、Y 的相關情形：



[解法]：

(A)完全正相關 (B)完全負相關 (C)零相關 (D)零相關  
(E)低度正相關 (F)完全曲線相關

(練習3) 下列有關兩變數  $X$  與  $Y$  的 8 個散布圖中



(1) 那些圖形較接近正相關？

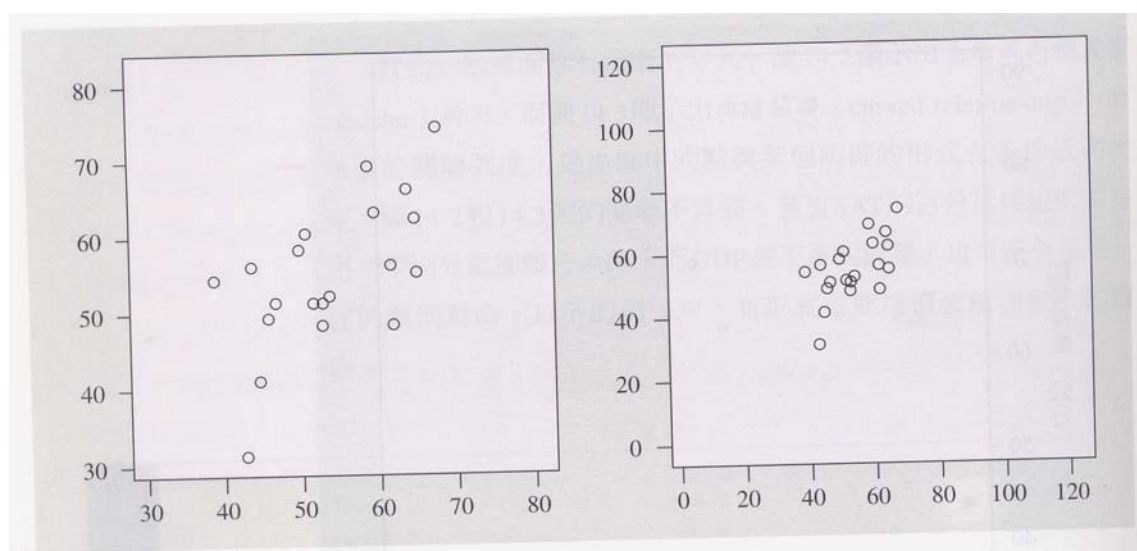
(2) 那些圖形較接近負相關？

(3) 那些圖形較接近零相關？

Ans : (1)(A)(E)(F)(G) (2)(C)(D)(G) (3)(B)

### (3) 相關係數(correlation)的引進

散布圖呈現兩個變數之間相關的方向、型式、強度。其中直線相關尤其重要，因為直線是最簡單的型態，但是光用眼睛看，並不容易判斷出相關的強度，如下圖，兩個散布圖畫的是同一組數據，只是兩個圖形的坐標選取之範圍不同，因此使得右圖看起來似乎有較強的直線相關。所以只要我們改一改散布圖上坐標軸的刻度或範圍，或是點和點之間的空白處大小，眼睛就可能受騙。所以得定義一個能夠衡量兩個變數直線相關強度的統計量，這就是**相關係數**。



(a)相關係數的定義：

衡量兩個變數直線相關的程度的統計量—相關係數定義如下：

對於兩組數據 X、Y

X	$x_1$	$x_2$	...	$x_n$
Y	$y_1$	$y_2$	...	$y_n$

定義相關係數

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

$$\text{其中 } x_i' = \frac{x_i - \bar{x}}{S_x}, y_i' = \frac{y_i - \bar{y}}{S_y} \text{ (標準化資料)}$$

[說明]：

根據之前的討論，我們知道  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  的正負表示相關程度的正負。然而當資料數據增加時，亦即樣本數  $n$  增加時，相對應的和  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  將隨之變大或變小，爲了消除這個影響的因素，將  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  加以規範化，而引進以上的定義。

(4)相關係數的性質：

(a)  $-1 \leq r \leq 1$

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

[代數的觀點]：

根據柯西不等式：

若設  $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n$  爲  $2n$  個實數，

$$\text{則 } (a_1^2 + a_2^2 + \dots + a_n^2)(b_1^2 + b_2^2 + \dots + b_n^2) \geq (a_1 b_1 + a_2 b_2 + \dots + a_n b_n)^2$$

將  $x_i - \bar{x}$  視爲  $a_i$ ， $y_i - \bar{y}$  視爲  $b_i$ ，即可得到  $r^2 \leq 1 \Leftrightarrow -1 \leq r \leq 1$

[向量的觀點]：

若令  $\vec{A} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$ ， $\vec{B} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$

$$\text{則 } r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| |\vec{B}|} = \cos \theta \Leftrightarrow -1 \leq r \leq 1$$

(此處的  $\cos \theta$  是一個抽象的概念)

(b)相關係數與單位無關：

若設 $x_i^* = a + bx_i$ ， $y_i^* = c + dy_i$ ， $i = 1, 2, \dots, n$ ，其中 $a, b, c, d$ 為給定之常數，則當 $bd > 0$ 時， $r = r^*$ ，當 $bd < 0$ 時， $r = -r^*$ 。

[證明]：

設 $\vec{A} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$ ， $\vec{B} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$

$\vec{A}^* = (x_1^* - \bar{x}^*, x_2^* - \bar{x}^*, \dots, x_n^* - \bar{x}^*)$ ， $\vec{B}^* = (y_1^* - \bar{y}^*, y_2^* - \bar{y}^*, \dots, y_n^* - \bar{y}^*)$

因為 $x_i^* = a + bx_i$ ， $y_i^* = c + dy_i$ ，所以 $\bar{x}^* = a + b\bar{x}$ ， $\bar{y}^* = c + d\bar{y}$

$\Rightarrow x_i^* - \bar{x}^* = b(x_i - \bar{x})$ ， $y_i^* - \bar{y}^* = d(y_i - \bar{y})$

$\Rightarrow \vec{A}^* = b\vec{A}$ ， $\vec{B}^* = d\vec{B}$

$\Rightarrow r^* = \frac{\vec{A}^* \cdot \vec{B}^*}{|\vec{A}^*| |\vec{B}^*|} = \frac{b\vec{A} \cdot d\vec{B}}{|b\vec{A}| |d\vec{B}|} = \left(\frac{bd}{|bd|}\right) \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| |\vec{B}|} = \left(\frac{bd}{|bd|}\right) \cdot r$

當 $bd > 0$ 時， $r = r^*$ ，當 $bd < 0$ 時， $r = -r^*$ 。

(c)正的 $r$ 值顯示變數之間有正相關，負的 $r$ 值顯示變數之間有負相關， $r$ 值若很接近 0，表示變數之間有很弱的直線相關。 $r = 1$ 時，表示樣本點都落在斜率為正的一條直線上， $r = -1$ 時，表示樣本點都落在斜率為負的一條直線上。

(d)相關係數會受少數極端觀測值得嚴重影響，如下圖，可以知道，極端值對相關係數的影響。

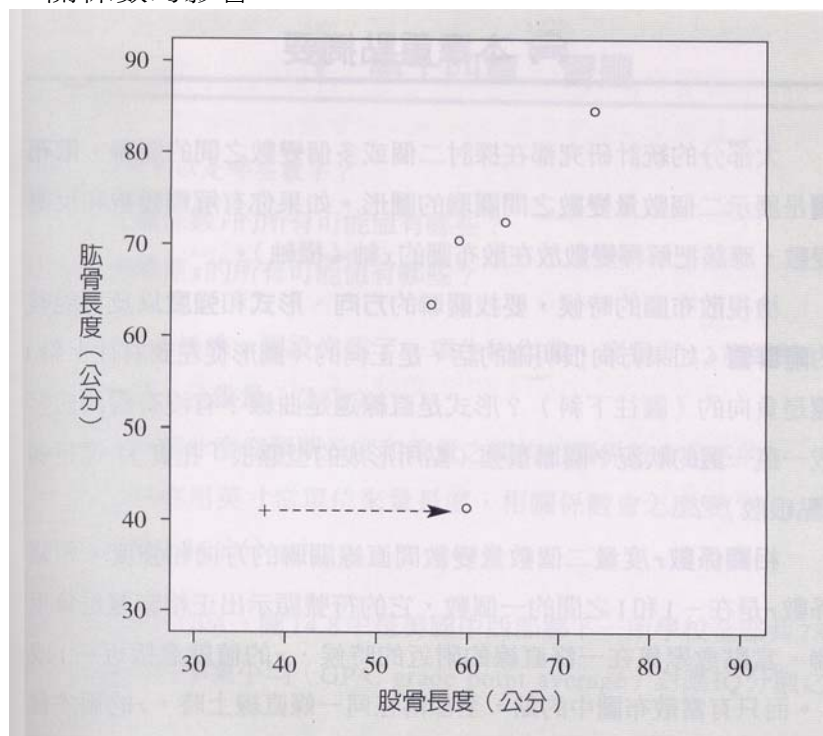


圖14.7 只移動一個點，就把相關係數從 $r = 0.994$ 變成 $r = 0.640$ 。

(e)兩個變數之間有很強的相關，也不一定代表兩者之間有因果關係。

例如：統計世界各國平均每人電視機數 $x$ 與人民的平均壽命 $y$ 。我們會得到很高的正相關，即有很多電視機的國家，人民的平均壽命較長。所謂的因果關係是

指只要改變 $x$ 的值，就可以使 $y$ 的值改變，換句話說，我們能否藉由運送一大堆的電視機來增加某些國家人民的平均壽命呢？當然不行！

富國的電視機比窮國多，而富國的人民平均壽命也比較長，但這是因為他們有比較好的營養、乾淨的飲水及較佳的醫療資源。電視機和壽命長短之間並沒有因果關係。

**[例題3]** 一肥皂廠商欲推出一種新產品，在上市之前以不同的單價  $x$ (單位:十元)，調查市場的需求量  $y$ (單位:萬盒)，調查結果如下：

$x$	8	9	10	11	12
$y$	11	12	10	8	9

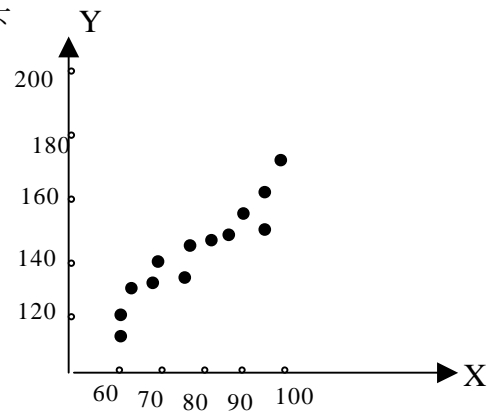
問  $x$ ， $y$  的相關係數最接近下列那一個值？ (84 學科)

(A)  $\frac{4}{5}$  (B)  $\frac{2}{5}$  (C) 0 (D)  $-\frac{2}{5}$  (E)  $-\frac{4}{5}$ 。 Ans : (E)

**[例題4]** 右圖為一班參加高中聯考成績， $X$  表示英文成績， $Y$  表示國文成績，兩個變數的相關係數最接近下列那一個值？

(A) 2 (B) 1 (C) 0.75 (D) 0.5 (E) 0.25

Ans : (C)



**[例題5]** 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？

(A)  $R_{wY}=7(24-R_{XY})$  (B)  $R_{wY}=7R_{XY}$  (C)  $R_{wY}=-7R_{XY}$   
 (D)  $R_{wY}=R_{XY}$  (E)  $R_{wY}=-R_{XY}$  Ans : (E) (90 學科)



(練習4)  $x, y$  平面上求樣本點(1,1)、(1,2)、(4,1)、(4,2)的相關係數  $r = ?$  Ans : 0

(練習5) 調查八位同學某次數學及物理抽考的成績為

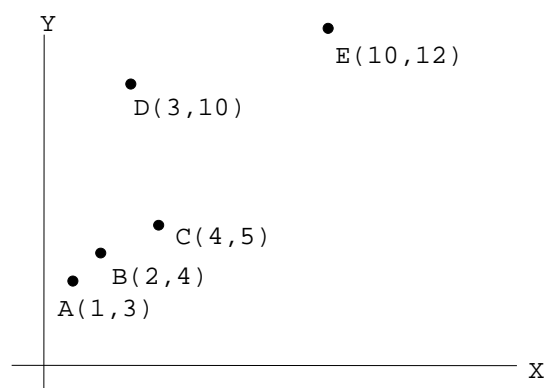
學生	A	B	C	D	E	F	G	H
數學	93	35	57	74	45	76	53	87
物理	73	37	54	70	54	82	48	62

試求其相關係數。 Ans : 0.82

(練習6) 如圖所示，有 5 筆(X,Y)資料。試問：  
去掉哪一筆資料後，剩下來 4 筆資料的相關係數最大？

(1)A (2)B (3)C (4)D (5)E

Ans : (4) (89.學科)



(練習7) 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：

學生	甲	乙	丙	丁	戊	己	庚	辛	壬	癸
缺課數	1	2	3	3	4	3	5	6	3	0
成績	100	90	90	80	70	70	60	60	80	100

設兩者的相關係數為  $r$ ，則

(A)  $-1 \leq r \leq -0.6$  (B)  $-0.6 < r < -0.2$  (C)  $-0.2 \leq r \leq 0.2$  (D)  $0.2 < r < 0.6$

(E)  $0.6 \leq r \leq 1$

Ans : (A) (86 自)

(練習8) 設  $X, Y$  的相關係數為  $r = 0.123$ ，且  $X' = -4X + 5$ ， $Y' = 6Y - 4$  的相關係數為  $r'$ ，則  $r' = ?$  Ans :  $r' = -0.123$

## (丙)最小平方方法

如果散布圖顯示出兩個數量變數之間的直線相關，我們會希望在散布圖中畫條直線，來對這個直線相關做一個概述。最小平方方法就是一種找出這樣的直線之方法，找出來的直線稱為最佳直線或迴歸直線，利用最佳直線可以利用一個變數來解釋或預測另一個變數，條件是它們之間的關係是可以解釋或預測的。

(1)最小平方方法：

例子：設樣本點 $(x_1, y_1)=(1, 2)$ 、 $(x_2, y_2)=(2, 1)$ 、 $(x_3, y_3)=(3, 3)$ ，求兩實數 $a, b$ 使得下列 $D$ 值最小： $D=(y_1-a-bx_1)^2+(y_2-a-bx_2)^2+(y_3-a-bx_3)^2$ 。

$$\therefore D = (2-a-b)^2 + (1-a-2b)^2 +$$

$$(3-a-3b)^2$$

$$= (a+b)^2 + (a+2b)^2 + (a+3b)^2$$

$$-4(a+b) - 2(a+2b) - 6(a+3b)$$

$$+4+1+9$$

$$= 3a^2 + 12ab + 14b^2 - 12a - 26b + 14$$

$$= 3(a^2 + 4ab + 4b^2) + 2b^2 - 12a - 26b$$

$$+14$$

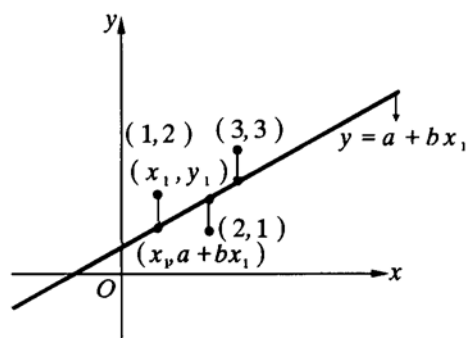
$$= 3(a+2b)^2 - 12(a+2b) + 2b^2 - 2b + 14$$

$$= 3[(a+2b)^2 - 4(a+2b) + 4] + 2(b^2 - b + \frac{1}{4}) + 2 - \frac{1}{2}$$

$$= 3(a+2b-2)^2 + 2(b-\frac{1}{2})^2 + \frac{3}{2}$$

$$\therefore \text{當} \begin{cases} a+2b-2=0 \\ b-\frac{1}{2}=0 \end{cases} \text{時 } D \text{ 的最小值為 } \frac{3}{2}$$

$$\text{即 } a=1, b=\frac{1}{2}$$



[幾何解釋]：

$D$ 的意義就是各樣本點與樣本點做直線的鉛直線的交點之距離平方和，所謂最佳直線 $L: y=a+bx$ 就是找到 $a, b$ ，使得 $D$ 的值最小。

最小平方方法：

對於給定有限個樣本點 $(x_1, y_1)$ 、 $(x_2, y_2)$ 、...、 $(x_n, y_n)$ ，要求出一條直線 $y=a+bx$ 使

得誤差的平方和 $E = \sum [y_i - (a + bx_i)]^2$ 最小。這樣的直線 $y=a+bx$ 稱為**最佳直線**或**迴**

**歸直線**。

(2)求最佳直線：

給定 $X$ 、 $Y$ 兩個變數，如表所示

$X$	$x_1$	$x_2$	$\cdots$	$x_n$
$Y$	$y_1$	$y_2$	$\cdots$	$y_n$

，欲找出 $a, b$ 使得誤差的平方

和 $E = \sum_{i=1}^n [y_i - (a + bx_i)]^2$ 最小。

[方法一]：

定義：

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2, \quad S_{YY} = \sum_{i=1}^n y_i^2 - n \cdot \bar{y}^2$$

$$\begin{aligned} S_{XY} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) \\ &= \sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x} \bar{y} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{y} \bar{x} + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \end{aligned}$$

$$(y_i - a - bx_i)^2 = a^2 + 2abx_i + b^2x_i^2 - 2ay_i - 2bx_iy_i + y_i^2$$

$$\begin{aligned} E &= \sum_{i=1}^n [y_i - (a + bx_i)]^2 \\ &= na^2 + 2ab \sum_{i=1}^n x_i + b^2 \sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n y_i - 2b \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 \\ &= na^2 + 2abn\bar{x} + b^2(S_{XX} + n\bar{x}^2) - 2an\bar{y} - 2b(S_{XY} + n\bar{x}\bar{y}) + S_{YY} + n\bar{y}^2 \\ &= n(a^2 + 2ab\bar{x} + b^2\bar{x}^2) + b^2S_{XX} - 2an\bar{y} - 2b(S_{XY} + n\bar{x}\bar{y}) + S_{YY} + n\bar{y}^2 \\ &= n(a + b\bar{x})^2 + b^2S_{XX} - 2n\bar{y}(a + b\bar{x}) - 2bS_{XY} + S_{YY} + n\bar{y}^2 \\ &= n[(a + b\bar{x})^2 - 2\bar{y}(a + b\bar{x}) + \bar{y}^2] + S_{XX}[b^2 - \frac{2S_{XY}}{S_{XX}}b + (\frac{S_{XY}}{S_{XX}})^2] - \frac{S_{XY}^2}{S_{XX}} + S_{YY} \\ &= n(a + b\bar{x} - \bar{y})^2 + S_{XX}(b - \frac{S_{XY}}{S_{XX}})^2 + (S_{YY} - \frac{S_{XY}^2}{S_{XX}}) \end{aligned}$$

$$\text{欲使 } E \text{ 的值最小} \Rightarrow \begin{cases} a + b\bar{x} - \bar{y} = 0 \\ b - \frac{S_{XY}}{S_{XX}} = 0 \end{cases} \Rightarrow b = \frac{S_{XY}}{S_{XX}}, \quad a = \bar{y} - \frac{S_{XY}}{S_{XX}} \bar{x}.$$

[方法二]：

將  $X$ 、 $Y$  兩個變數標準化化成  $X'$ 、 $Y'$ ，其中  $X' = \frac{X - \bar{x}}{S_X}$ ， $Y' = \frac{Y - \bar{y}}{S_Y}$

$X'$ 、 $Y'$  的平均數與標準差分別為 0 與 1

標準化後，設最佳直線  $L'$ ： $y' = a + bx'$

$$\begin{aligned} \text{誤差的平方和 } E' &= \sum_{i=1}^n [y'_i - (a + bx'_i)]^2 \\ &= \sum_{i=1}^n [(y'_i)^2 - 2y'_i(a + bx'_i) + (a + bx'_i)^2] \\ &= \sum_{i=1}^n (y'_i)^2 - 2 \sum_{i=1}^n (ay'_i + bx'_i y'_i) + \sum_{i=1}^n (a^2 + 2abx'_i + b^2(x'_i)^2) \\ &= \sum_{i=1}^n (y'_i)^2 - 2a \sum_{i=1}^n y'_i - 2b \sum_{i=1}^n x'_i y'_i + na^2 + 2ab \sum_{i=1}^n x'_i + b^2 \sum_{i=1}^n (x'_i)^2 \\ &\quad \text{因爲 } X'、Y' \text{ 的平均數與標準差分別為 0 與 1} \end{aligned}$$

$$\begin{aligned}
& \text{所以 } \sum_{i=1}^n y'_i = \sum_{i=1}^n x'_i = 0 \\
& = \sum_{i=1}^n (y'_i)^2 - 2b \sum_{i=1}^n x'_i y'_i + na^2 + b^2 \sum_{i=1}^n (x'_i)^2 \\
& = \sum_{i=1}^n (y'_i)^2 + na^2 + \sum_{i=1}^n (x'_i)^2 \left[ b - \frac{\sum_{i=1}^n x'_i y'_i}{\sum_{i=1}^n (x'_i)^2} \right]^2 - \frac{(\sum_{i=1}^n x'_i y'_i)^2}{\sum_{i=1}^n (x'_i)^2}
\end{aligned}$$

當  $a=0$ ， $b = \frac{\sum_{i=1}^n x'_i y'_i}{\sum_{i=1}^n (x'_i)^2}$  時， $E'$  的值最小。

$$\begin{aligned}
& \text{另一方面，} b = \frac{\sum_{i=1}^n x'_i y'_i}{\sum_{i=1}^n (x'_i)^2} = \frac{\sum_{i=1}^n (\frac{x_i - \bar{x}}{S_X})(\frac{y_i - \bar{y}}{S_Y})}{\sum_{i=1}^n (\frac{x_i - \bar{x}}{S_X})^2} \\
& = \frac{S_X^2}{S_X S_Y} \left( \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
& = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} = \text{相關係數 } r。
\end{aligned}$$

$\Rightarrow$  最佳直線  $y' = rx'$ 。

再將  $x' = \frac{x - \bar{x}}{S_X}$ ， $y' = \frac{y - \bar{y}}{S_Y}$  代入上式

$$\Rightarrow \frac{y - \bar{y}}{S_Y} = r \left( \frac{x - \bar{x}}{S_X} \right) \Rightarrow y - \bar{y} = \frac{r S_Y}{S_X} (x - \bar{x})$$

最佳直線  $L: y = a + bx$ ，其中  $b = \frac{r S_Y}{S_X} = \frac{S_{XY}}{S_{XX}}$ ， $a = \bar{y} - \frac{S_{XY}}{S_{XX}} \bar{x}$

由上式可知最佳直線必過點  $(\bar{x}, \bar{y})$ 。

結論：

(1) 給定  $X$ 、 $Y$  兩個變數，如表所示  $\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ Y & y_1 & y_2 & \cdots & y_n \end{array}$ ，將  $X$ 、 $Y$  兩個變數標準化，化成  $X'$ 、 $Y'$ ，最佳直線  $L'$  為  $y' = rx'$ ，其中  $r$  為  $X$ 、 $Y$  的相關係數。

(2) 若給定  $X$ 、 $Y$  兩個變數，如表所示  $\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ Y & y_1 & y_2 & \cdots & y_n \end{array}$ ，

則  $Y$  對  $X$  的最佳直線  $L: y = a + bx$  必通過點  $(\bar{x}, \bar{y})$ ，

其中  $b = \frac{r S_Y}{S_X} = \frac{S_{XY}}{S_{XX}}$ ， $a = \bar{y} - \frac{S_{XY}}{S_{XX}} \bar{x}$ 。

[例題6] 高三某班有 10 位同學(編號 1,2,...,10)，其期末考成績與該學期上課時缺課數的統計資料如下：

編號	1	2	3	4	5	6	7	8	9	10
缺課數	0	3	6	5	3	4	3	3	2	1
成績	95	75	55	55	65	65	75	85	85	95

- (1)試求這 10 個學生的缺課數  $X$  與期末成績  $Y$  的相關係數。  
 (2)求這 10 個資料變數  $Y$  對變數  $X$  的最佳直線方程式。  
 (3)根據這條最佳直線，請預測缺課數為 7 時的成績為多少？

Ans：(1)-0.93 (2) $y=75-7.86(x-3)$  (3)43.56

[例題7] 設有一隨機樣本包含 200 對父子體重( $x_i, y_i$ )的觀察資料，且已算出下列的統計

量(單位為公斤)： $\bar{x}=68$ ， $\bar{y}=69$ ， $\sum_{i=1}^{200} (x_i - \bar{x})^2 = 1920$ ， $\sum_{i=1}^{200} (y_i - \bar{y})^2 = 2040$

$$\sum_{i=1}^{200} (x_i - \bar{x})(y_i - \bar{y}) = 1920,$$

試求(1)兩變數 $X$ 與 $Y$ 的相關係數。

(2)求變數 $Y$ 對 $X$ 的最佳直線方程式。

Ans：(1)0.51 (2) $y=69+0.53(x-68)$

(練習9) 設抽樣某班 8 位學生的數學成績(x)與英文成績(y)，結果如下

$$\bar{x}=65, \bar{y}=70, S_X=10, S_Y=5, r=0.8$$

(1)請寫出英文成績(y)對數學成績(x)的迴歸式。

(2)若此班某位同學數學成績 65 分，請預測此生的英文成績。

Ans：(1) $y=44+0.4x$  (2)70 分

(練習10) 蒐集台灣地區 8 個地點的公告地價與市價(單位：萬元/坪)如下：

公告地價(x)	12	10	22	30	8	40	20	18
市價(y)	15	11	28	40	10	72	39	25

(1)試畫市價對公告地價的散布圖。

(2)試求市價對公告地價的相關係數。

(3)試求市價對公告地價的迴歸式。

(4)若某塊土地公告地價是每坪 28 萬元，試利用上面的迴歸式預測其市價。

Ans：(2)0.9626 (3) $y=6.6667+1.8333x$  (4)57.9991 萬

(練習11) 設某公司隨機抽樣 10 位員工的年齡(x)與血壓(y)的資料，結果算出

$$\sum_{i=1}^{10} x_i = 450, \sum_{i=1}^{10} y_i = 1300, \sum_{i=1}^{10} x_i^2 = 21250, \sum_{i=1}^{10} y_i^2 = 171250, \sum_{i=1}^{10} x_i y_i = 59100$$

(1)請問年齡與血壓的相關係數=？

(2)請寫出血壓對年齡的最佳直線方程式。

(3)此公司員工的年齡 50 歲，請預測此員工的血壓是多少？

Ans：(1)0.4 (2) $y=103+0.6x$  (3)133

### 綜合練習

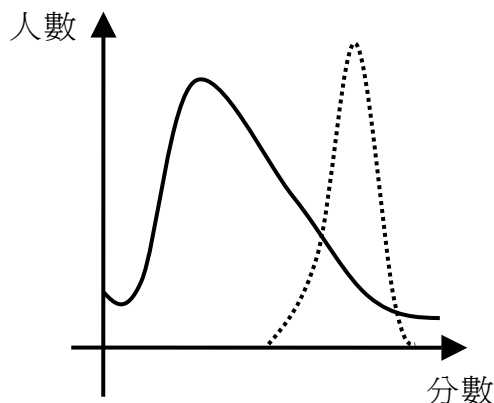
- (1) 某班數學老師算出學生學習成績後，鑒於學生平時都很用功，決定每人各加 5 分(加分後沒人超過滿分)，則加分前與加分後，學生成績統計數值絕對不會改變的有(A)算術平均數 (B)中位數 (C)標準差 (D)變異係數 (E)全距

(88 自)

- (2) 某年聯考甲乙兩科成績的直方圖如圖所示，(由於考生人數眾多，成績分佈的直方圖可視為平滑的曲線)，則下列那些敘述是正確的？

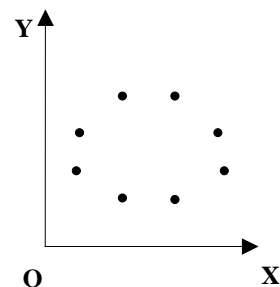
- (A)甲的算術平均數比乙的算術平均數大  
(B)甲的中位數比乙的中位數大  
(C)甲的全距比乙的全距大  
(D)甲的標準差比乙的標準差大  
(E)甲的變異係數比乙的變異係數大

(87 自)



- (3) 右圖表兩組數據  $x, y$  的分佈圖，試問其相關係數  $r$  最接近下列何值？

- (A)1 (B)0.5 (C)0 (D)-0.5 (E)-1 (88 社)



- (4) 某班的 50 名學生參加一項考試，考題共 100 題，全為 5 選 1 的單選題。計分法共有  $X, Y$  兩種：某學生有  $N$  題放棄沒答， $R$  題答對， $W$  題答錯，則  $X=R-\frac{W}{4}$ ，

$Y=R+\frac{N}{5}$ ，試問下列敘述那些是正確的？

- (A)同一班學生的  $X$  分數不可能大於  $Y$  分數。  
(B)全班  $X$  分數的算術平均數不可能大於  $Y$  分數的算術平均數。  
(C)任兩學生  $X$  的分數差之絕對值不可能大於  $Y$  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。 (90 自)

- (5) 假設某班有 40 人，最近兩次數學測驗每一位同學第一次成績都比第二次少 8 分，那麼下列有關這兩次數學測驗成績的統計結果哪一個是錯誤的？

- (A)全距相等(B)算術平均數相等(C)四分位差相等(D)標準差相等(E)正相關

- (6) 數學老師想把某次模擬考滿分 120 分的成績( $X$ )作調整為滿分 100 分的平時成績( $Y$ )，以便登記成一次平時成績，故  $Y=\frac{5}{6}X$ 。現在模擬考的成績求得算術平均數  $\bar{x}$ ，中位數  $Me$ ，全距  $D$ ，標準差  $S$ ，數學與物理分數相關係數為  $r$ ；若調整之後，各相對統計量為算術平均數  $\bar{x}'$ ，中位數  $Me'$ ，全距  $D'$ ，標準差  $S'$ ，數學與物

理分數相關係數為 $r'$ ，則下列何者正確？(A)  $\bar{x}' = \frac{5}{6}x$  (B)  $Me = Me'$  (C)  $D' = \frac{5}{6}D$   
 (D)  $S' = \frac{5}{6}S$  (E)  $r = r'$ 。

(7) 十位考生之國文與數學成績列表如下：

考生編號	1	2	3	4	5	6	7	8	9	10
國文	89	65	76	69	82	57	66	72	78	66
數學	75	57	65	65	83	63	58	62	63	69

今已算出國文成績之標準差為 8.9(取至小數點第一位)，數學成績之標準差為 7.5(取至小數點第一位)，

(a) 設國文成績的平均為 $\bar{X}$ ，而變異係數為 $CV_1$ ，數學成績的平均為 $\bar{Y}$ ，而變異係數為 $CV_2$ ，則

(A)  $CV_1 = CV_2$  (B)  $CV_1 < CV_2$  (C)  $\bar{X} = 72$  (D)  $\bar{Y} = 64$  (E)  $\bar{X} = \bar{Y}$

(b) 則此十位考生兩科成績之相關係數最接近

(A) -0.85 (B) 0.25 (C) 0.66 (D) 0.78 (E) 0.85

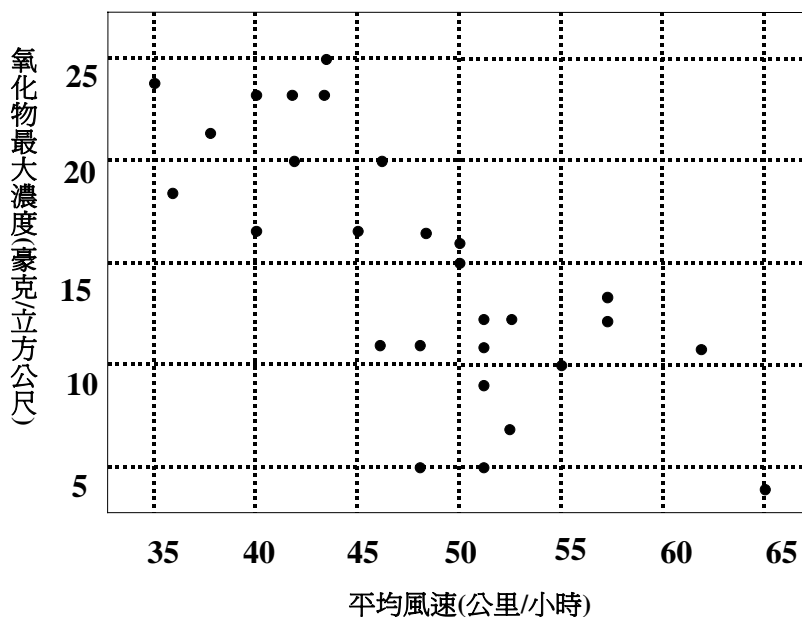
(8) 空氣品質會受到污染物排放量及大氣擴散等因素的影響。某一機構為了解一特定地區的空氣品質，連續二十八天蒐集了該地區早上的平均風速及空氣中某特定氧化物的最大濃度。再繪製這二十八筆資料的散佈圖(見下圖)，現根據該圖，可知

(A) 此筆資料，該氧化物最大濃度的標準差大於 15。

(B) 此筆資料，該氧化物最大濃度的中位數為 15。

(C) 此筆資料，平均風速的中位數介於 45 與 50 之間。

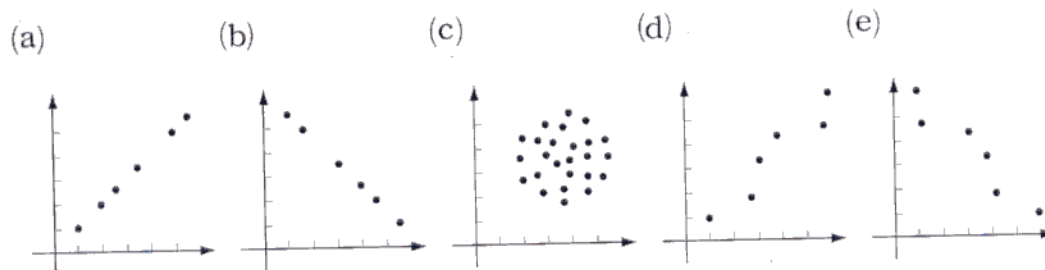
(D) 若以最小平方方法決定數據集中直線趨勢的直線，則該直線的斜率小於 0。



(91 指定甲)



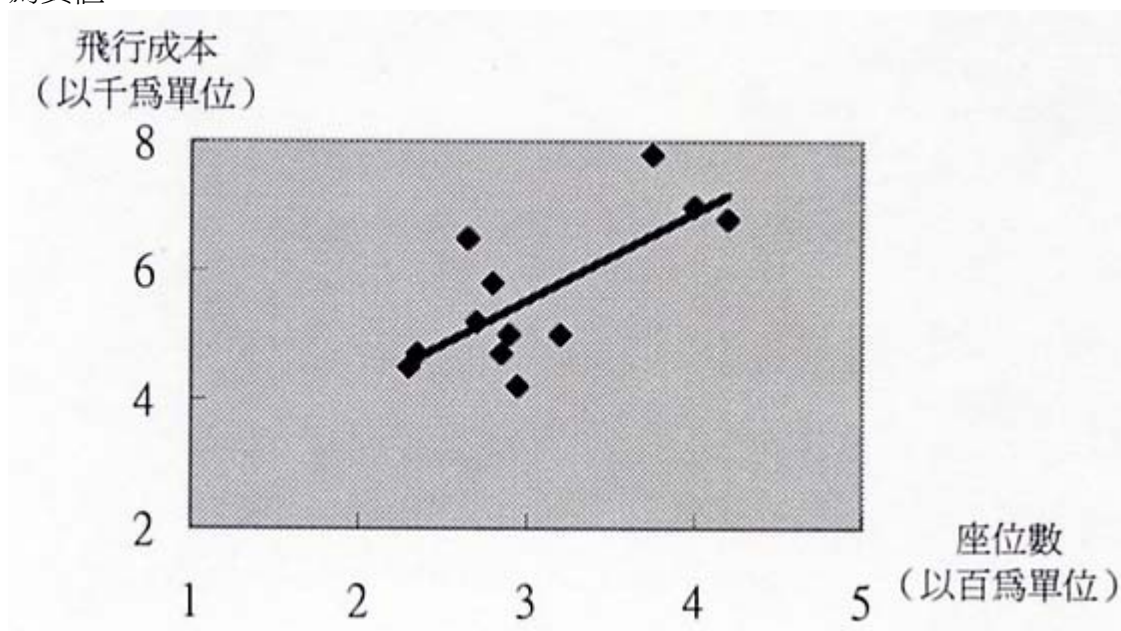
(9) 請排出下面 5 個散佈圖中  $x, y$  的相關係數的大小順序。



(10) 右圖是一個航空公司旗下 12 架飛機的散佈圖，橫軸座位數，縱軸為單位小時飛行成本，

圖中的直線代表數據集中趨勢的直線。請問下列敘述那些是正確的？

- (A) 根據此圖可知飛機座位數與單位小時飛行成本二者之間的相關係數為負。  
 (B) 圖中的直線斜率介於 10 與 20 間。  
 (C) 單就座位數介於 250 及 350 間七架中型飛機而言，圖中直線仍可視為代表該筆數據集中趨勢的直線。  
 (D) 單就座位數介於 250 及 350 間七架中型飛機而言，代表該筆數據集中趨勢的直線斜率為負值。



(11) 如下表，為 9 位同學參加大考中心舉辦的學科能力測驗數學科成績，其中有兩位同學不願透露成績，但由老師所有資料可知，9 位同學的平均成績為 12 分，

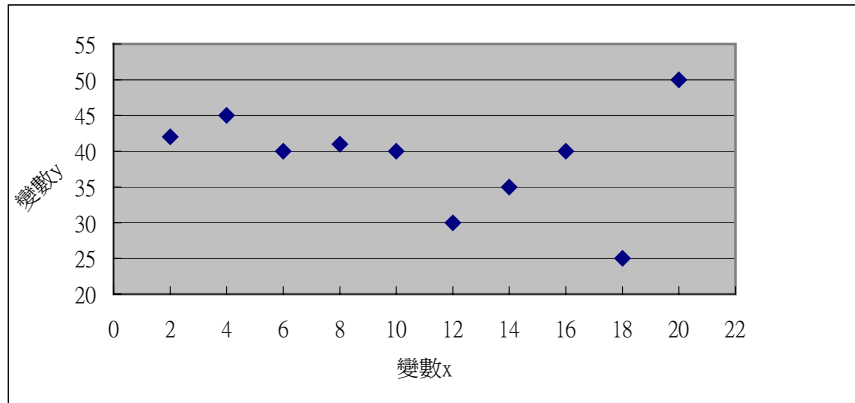
變異係數為  $\frac{50\sqrt{3}}{9}\%$ ，且已知 4 號同學的成績較 7 號同學好，求  $x, y$ ?

座號	1	2	3	4	5	6	7	8	9
成績	11	12	11	$x$	12	13	$y$	12	13

(12) 某種新藥的用量  $X$ (毫克數)與藥效期間  $Y$ (天數)的關係，經調查後資料如右表：

- (a) 若  $X$  與  $Y$  的相關係數為  $\sqrt{\frac{13}{k}}$ ，則  $k=$ \_\_\_\_\_。
- (b)  $Y$  對  $X$  的迴歸直線為  $y=ax+b$ ，則數對  $(a,b)=$ \_\_\_\_\_。
- (c) 若醫生用藥量為 12 毫克，試用迴歸直線預估藥效可達\_\_\_\_\_天。

- (13) 右圖為兩變數  $(x,y)$  10 筆資料的散布圖，  
試問下列敘述那些是正確的？



$$(\text{標準差 } S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}})$$

- (A) 變數  $y$  的中位數為 40
- (B) 變數  $y$  的平均數不大於 50
- (C) 變數  $y$  的標準差不大於 12
- (D) 變數  $x$  與變數  $y$  為正相關
- (E) 變數  $x, y$  的最佳直線斜率為負的

## 綜合練習解答

- (1) (C)(E)
- (2) (C)(D)(E)
- (3) (C)
- (4) (A)(B)(D)(E)
- (5) (B)
- (6) (A)(D)(E)
- (7) (a)C (b)C
- (8) (C)(D)
- (9) (a)>(d)>(c)>(e)>(b)
- (10) (B)(D)
- (11)  $x=14, y=10$
- (12) (a)14 (b)(2,3) (c)27
- (13) (A)(B)(C)(E)