第三十八單元 雙變量數據分析

(甲)相關係數

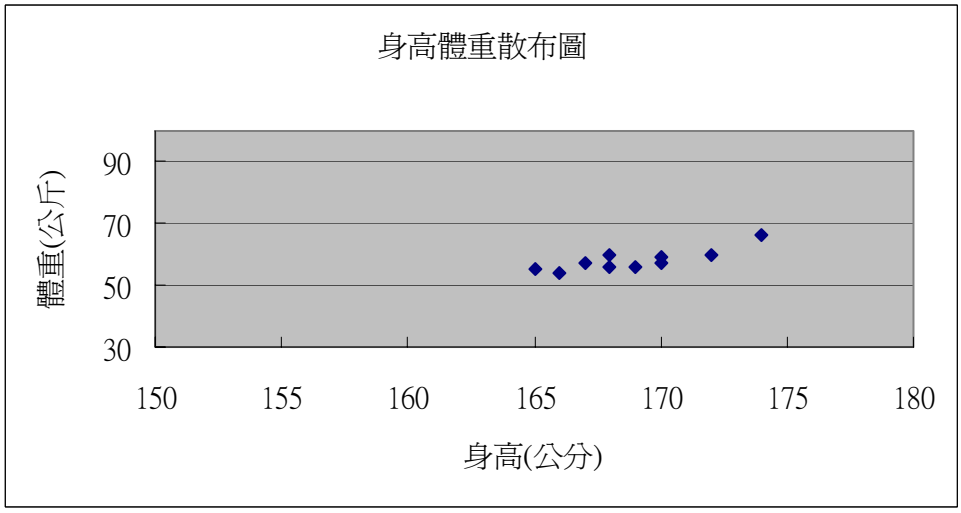
通常身高很高的人，體重不會太輕，物理成績高的學生，數學成績通常也不會很低，不管是身高、體重或是物理、數學成績，我們如何來衡量這兩個資料間的關係呢？可否由身高來預測體重，或是由數學成績來預測物理成績呢？

(1)散布圖(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，稱爲**散布圖**，散布圖上的點稱爲**樣本點**。



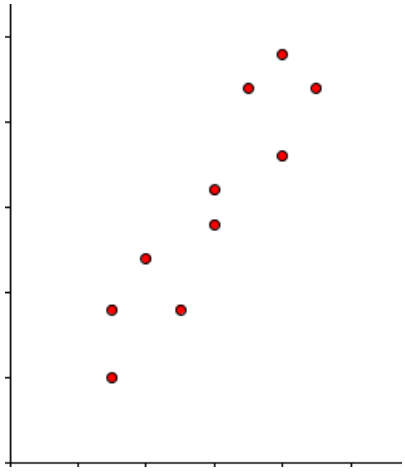
(2)散布圖與相關程度：

某種新藥的用量 X(毫克數)與藥效期間 Y(天數)的關係，

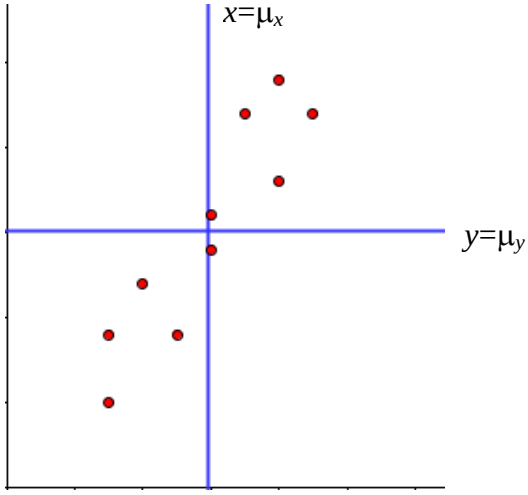
經調查後得到資料如下表：

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μ_x=5.9 毫克，藥效期間平均值μ_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爲增或同時爲減，兩者之間是有某種程度的相關性。

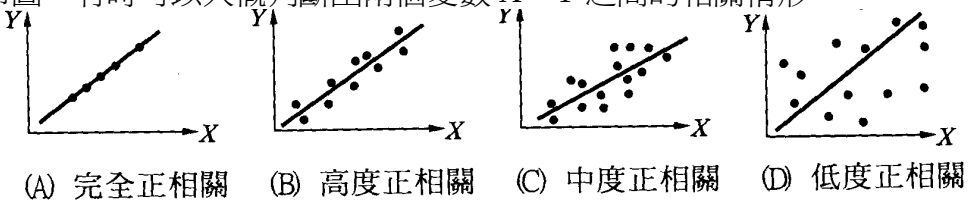
一般而言，如果在散布圖中以 $y=\mu_y$ 爲新的橫軸，$x=\mu_x$ 爲新的縱軸，則可將全圖分成四個象限，在第一三象限內的點(x_i,y_i)，$(x_i-\mu_x)(y_i-\mu_y)$的值爲正，；在第二四象限內的點(x_i,y_i)，$(x_i-\mu_x)(y_i-\mu_y)$的值爲負，若資料內的樣本點(x_1,y_1)、(x_2,y_2)、…、(x_n,y_n)中，計算$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$的值：

(a)若$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)>0$，則表示 X 與 Y 的變動趨勢大致相同，即同時爲增或同時爲減，我們稱兩者爲**正相關**。

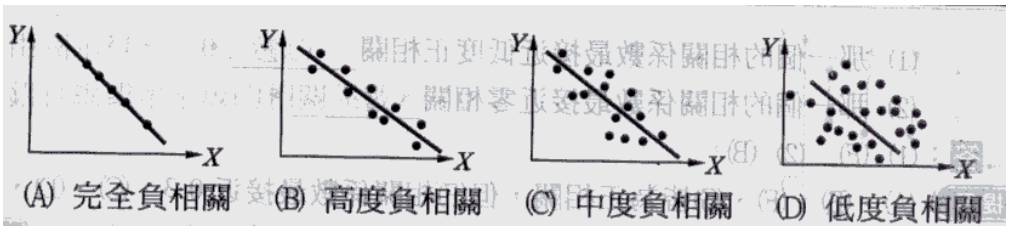
(b)若$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)<0$，則表示 X 與 Y 的變動趨勢大致相反，即此增彼減或此減彼增，我們稱兩者爲**負相關**。

(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：

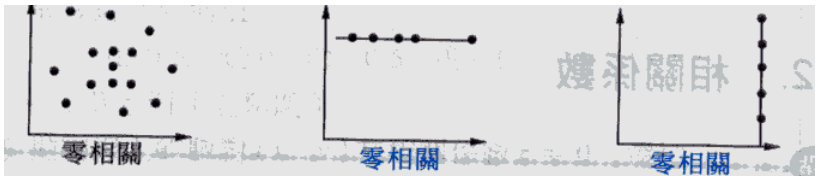
正相關：



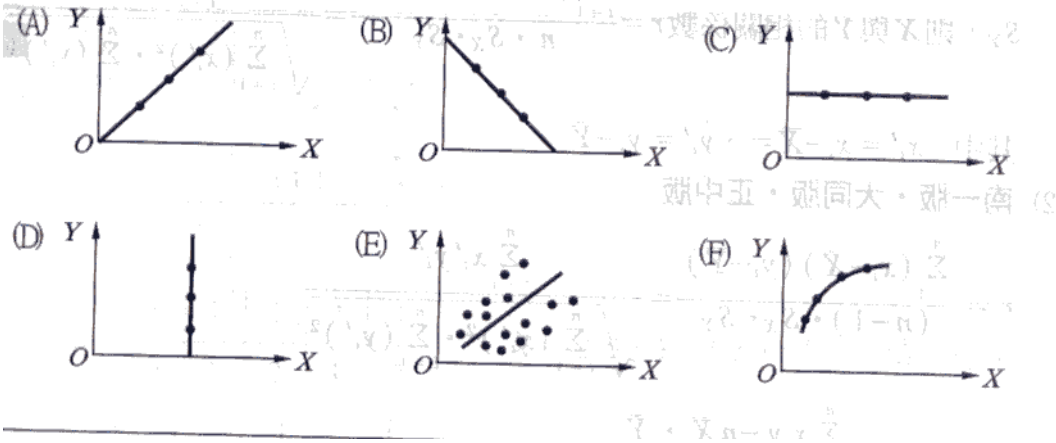
負相關：



零相關：

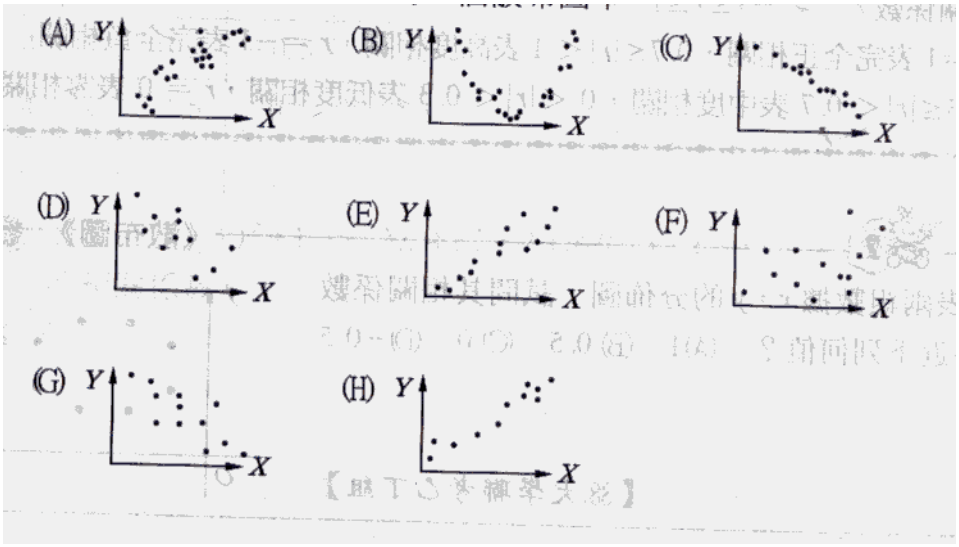


[例題1] 就下列圖形說明變量 X、Y 的相關情形：



[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1) 下列有關兩變數 X 與 Y 的 8 個散布圖中

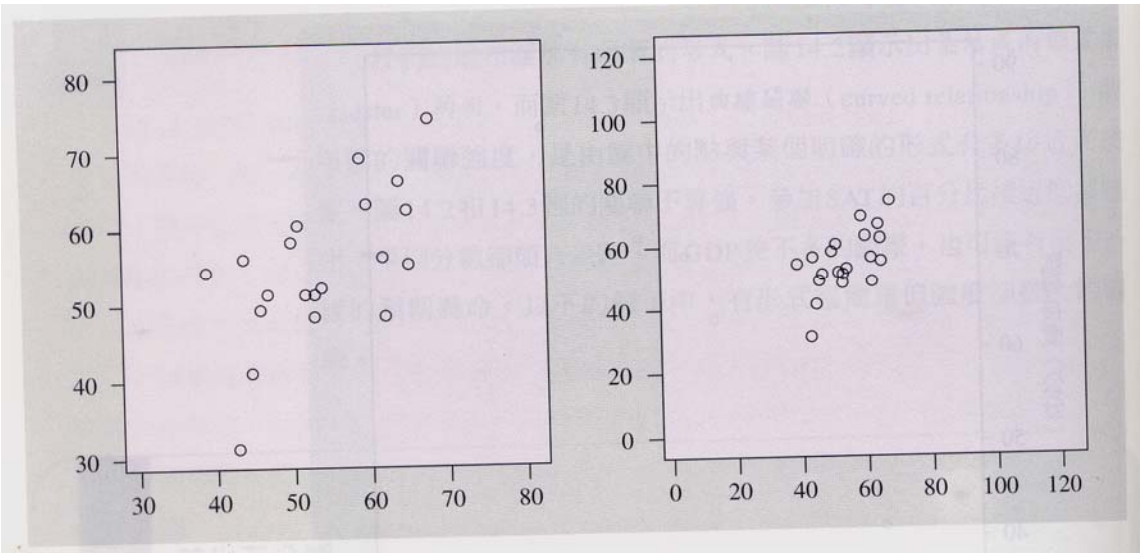


(1)那些圖形較接近正相關？
(2)那些圖形較接近負相關？
Ans：(1)(A)(E)(F)(H)　(2)(C)(D)(G)

(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，因爲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

$$r=\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot\sum_{i=1}^{n}(y_i-\mu_y)^2}}=\frac{\sum_{i=1}^{n}x_iy_i-n\cdot\mu_x\cdot\mu_y}{n\cdot\sigma_x\cdot\sigma_y}=\frac{\sum_{i=1}^{n}x_i^{/}\cdot y_i^{/}}{n}\ ,$$

$x_i^{/}=\frac{x_i-\mu_x}{\sigma_x}$ ，$y_i^{/}=\frac{y_i-\mu_y}{\sigma_y}$ (標準化資料)

其中μ_x、μ_y為 X、Y 的算術平均數；σ_x、σ_y為 X、Y 的標準差。

[說明]：

根據之前的討論，我們知道$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$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$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$將隨之變大或變小，爲了消除這個影響的因素，將$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$加以規範化，而引進以上的定義。

(4)相關係數的性質：

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

相關係數 $r=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot\sum_{i=1}^{n}(y_i-\mu_y)^2}}$ $\Rightarrow r^2=\dfrac{(\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y))^2}{\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot\sum_{i=1}^{n}(y_i-\mu_y)^2}$

[代數的觀點]：

根據柯西不等式：

若設 a_1、a_2、…、a_n，b_1、b_2、…、b_n 爲 $2n$ 個實數，

則$(a_1^2+a_2^2+\ldots+a_n^2)(b_1^2+b_2^2+\ldots b_n^2)\geq(a_1b_1+a_2b_2+\ldots+a_nb_n)^2$

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 $\overrightarrow{A}=(x_1-\mu_x,x_2-\mu_x,\ldots,x_n-\mu_x)$，$\overrightarrow{B}=(y_1-\mu_y,y_2-\mu_y,\ldots,y_n-\mu_y)$

則 $r=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot\sum_{i=1}^{n}(y_i-\mu_y)^2}}=\dfrac{\overrightarrow{A}\cdot\overrightarrow{B}}{|\overrightarrow{A}||\overrightarrow{B}|}=\cos\theta \Leftrightarrow -1\leq r\leq 1$

(此處的 cosθ是一個抽象的概念)

(b)相關係數的絕對值與單位無關：

若設 $x_i^*=a+bx_i$，$y_i^*=c+dy_i$，$i=1,2,\ldots,n$，其中 a,b,c,d 爲給定之常數，

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

[證明]：

設 $\overrightarrow{A}=(x_1-\mu_x,x_2-\mu_x,\ldots,x_n-\mu_x)$，$\overrightarrow{B}=(y_1-\mu_y,y_2-\mu_y,\ldots,y_n-\mu_y)$

$\overrightarrow{A^*}=(x_1^*-\mu^*_x,x_2^*-\mu^*_x,\ldots,x_n^*-\mu^*_x)$，$\overrightarrow{B^*}=(y_1^*-\mu^*_y,y_2-\mu^*_y,\ldots,y_n-\mu^*_y)$

因爲 $x_i^*=a+bx_i$，$y_i^*=c+dy_i$，所以$\mu^*_x=a+b\mu_x$，$\mu^*_y=c+d\mu_y$

$\Rightarrow x_i^*-\mu^*_x=b(x_i-\mu_x)$，$y_i^*-\mu^*_y=d(y_i-\mu_y)$

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

$\Rightarrow r^*=\dfrac{\overrightarrow{A^*}\cdot\overrightarrow{B^*}}{|\overrightarrow{A^*}||\overrightarrow{B^*}|}=\dfrac{b\overrightarrow{A}\cdot d\overrightarrow{B}}{|b\overrightarrow{A}||d\overrightarrow{B}|}=(\dfrac{bd}{|bd|})\dfrac{\overrightarrow{A}\cdot\overrightarrow{B}}{|\overrightarrow{A}||\overrightarrow{B}|}=(\dfrac{bd}{|bd|})\cdot r$

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爲正的一條直線上，$r=-1$ 時，表示樣本點都落在斜率爲負的一條直線上。

(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。

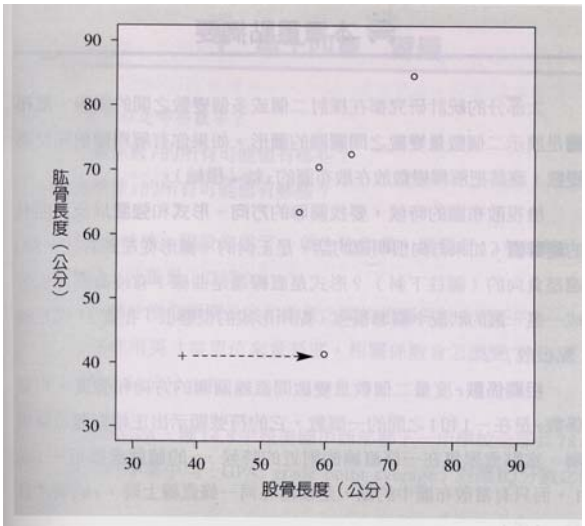


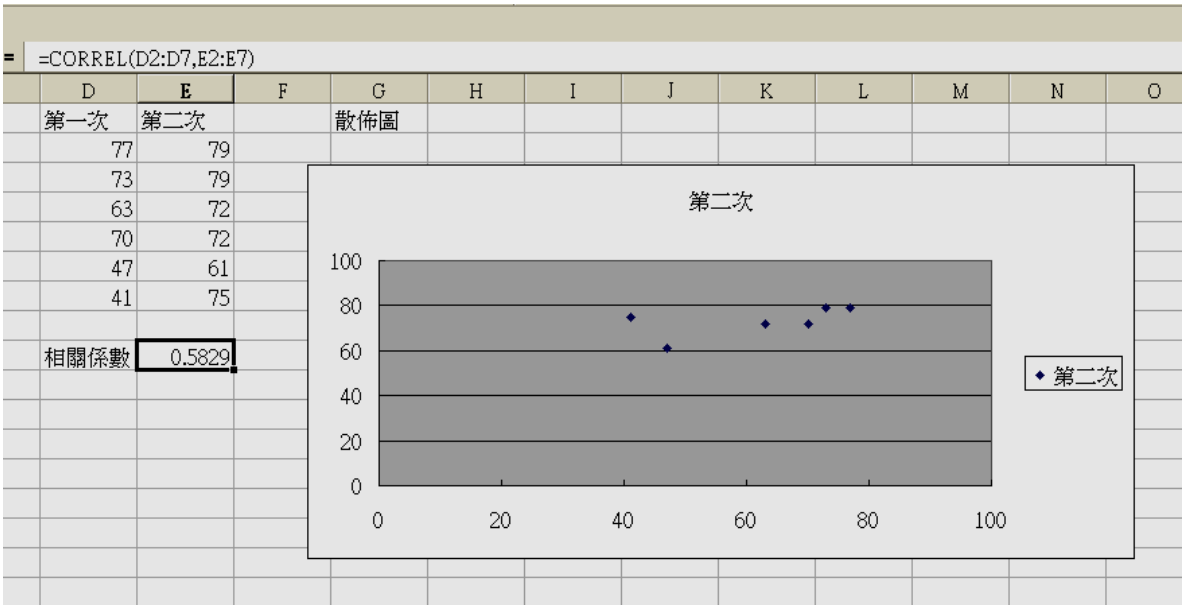
圖14.7 只移動一個點，就把相關係數從$r = 0.994$變成$r = 0.640$。

(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，但這是因爲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。

(5)用 Excel 計算相關係數：



[例題2] 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3] 調查某國某一年 5 個地區的香煙與肺癌之相關性，所得的數據爲(x_i, y_i)，i=1,2,3,4,5，其中變數 X 代表每人每年香煙消費量(單位：十包)，Y 代表每十萬人死於肺癌的人數。若已計算出下列數值：

$\sum_{i=1}^{5} x_i$ =135，$\sum_{i=1}^{5} x_i^2$ =3661，$\sum_{i=1}^{5} x_i y_i$ =2842，$\sum_{i=1}^{5} y_i$ =105，$\sum_{i=1}^{5} y_i^2$ =2209，

則 X 與 Y 的相關係數 r=_____。(2010 指定乙)

(參考說明：相關係數

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

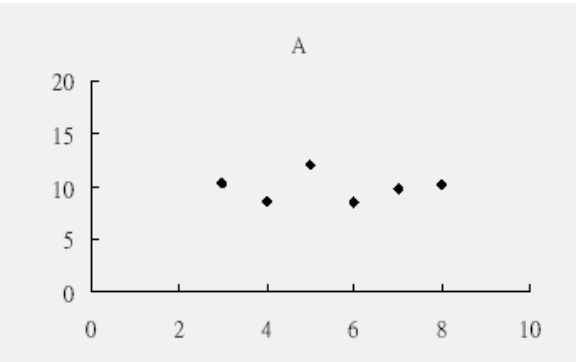
[答案]：0.875

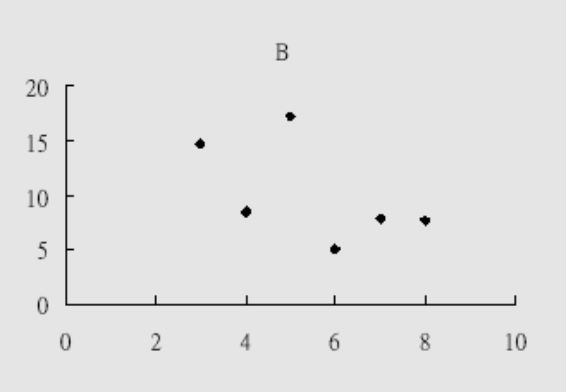
[解答]：

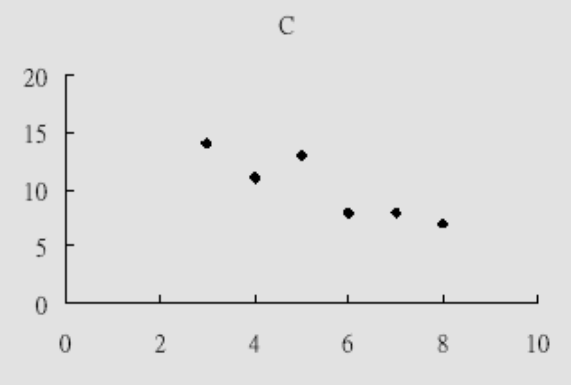
$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\cdot\bar{x}\cdot\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n\cdot\bar{x}^2}\cdot\sqrt{\sum_{i=1}^{n}y_i^2-n\cdot\bar{y}^2}}=\frac{2842-5\times27\times21}{\sqrt{3661-5(27)^2}\cdot\sqrt{2209-5(21)^2}}$$

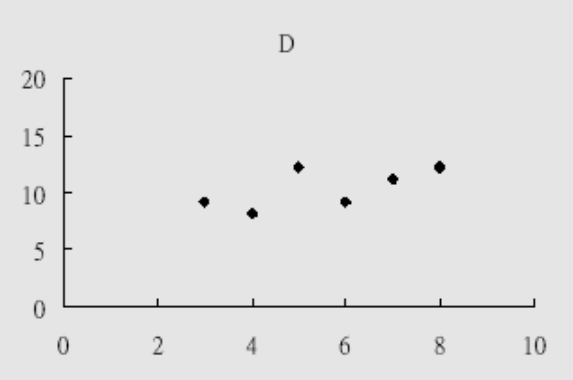
$=\frac{7}{\sqrt{16}\sqrt{4}}=\frac{7}{8}$ =0.875。

[例題4] A,B,C,D 是四組資料的散佈圖，如圖所示。利用最小平方法計算它們的迴歸直線， 發現有兩組資料的迴歸直線相同， 試問是哪兩組？
(1) A、B (2) A、C (3) A、D (4) B、C (5) B、D (2009 指定乙)









Ans：(4)

[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 之相關係數爲 R_{XY}，W,Y 之相關係數爲 R_{WY}，則 R_{XY} 與 R_{WY} 兩數之間的關係，下列選項何者爲真？
(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2) 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3) 調查八位同學某次數學及物理抽考的成績爲

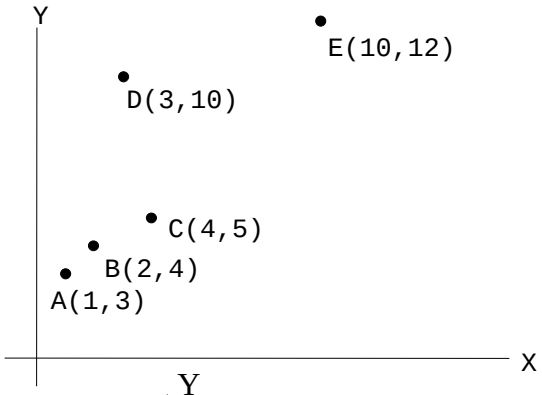
學生	A	B	C	D	E	F	G	H
數學	93	35	57	74	45	76	53	87
物理	73	37	54	70	54	82	48	62

試求其相關係數。　Ans：0.82

(練習4) 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

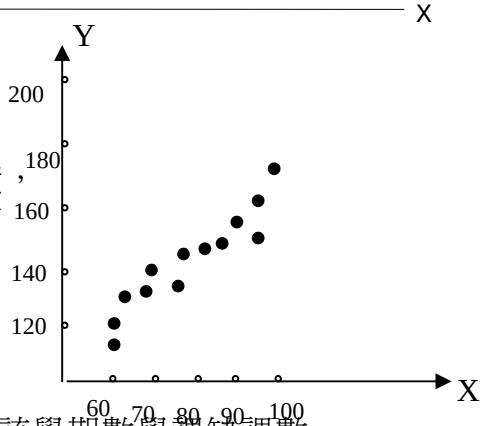
Ans：(4)　(89.學科)



(練習5) 右圖爲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)



(練習6) 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

設兩者的相關係數爲 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) $0.6 \leqq r \leqq 1$　　Ans：(*A*)

(練習7) 設 X、Y 的相關係數爲 r=0.123，且 $X'=-4X+5$，$Y'=6Y-4$ 的相關係數爲 r'，則 r'=？　Ans：$r'=-0.123$

(丙)最小平方法

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。最小平方法就是一種找出這樣的直線之方法，找出來的直線稱爲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：$D=(y_1-a-bx_1)^2+(y_2-a-bx_2)^2+(y_3-a-bx_3)^2$。

$\because D=(2-a-b)^2+(1-a-2b)^2+(3-a-3b)^2$

$=(a+b)^2+(a+2b)^2+(a+3b)^2-4(a+b)-2(a+2b)-6(a+3b)+4+1+9$

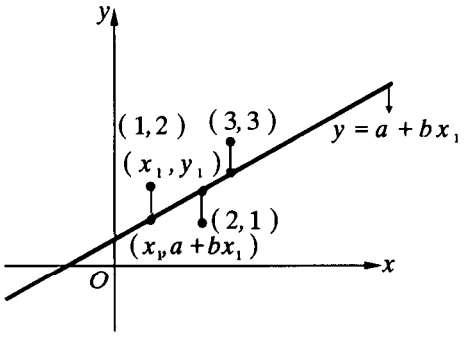
$=3a^2+12ab+14b^2-12a-26b+14$

$=3(a^2+4ab+4b^2)+2b^2-12a-26b+14$

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

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

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



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

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 y 對 x 的最佳直線 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 $y=a+bx$ 稱爲 y 對 x 的**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 $\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ \hline Y & y_1 & y_2 & \cdots & y_n \end{array}$ ，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：

$S_{XX}=\sum_{i=1}^{n}(x_i-\mu_x)^2$ ，$S_{YY}=\sum_{i=1}^{n}(y_i-\mu_y)^2$ ，$S_{XY}=\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$

$S_{XX}=\sum_{i=1}^{n}(x_i-\mu_x)^2=\sum_{i=1}^{n}x_i^{\,2}-n\cdot\mu_x$ ，$S_{YY}=\sum_{i=1}^{n}y_i^{\,2}-n\cdot\mu_y$

$S_{XY}=\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)=\sum_{i=1}^{n}(x_iy_i-\mu_xy_i-x_i\mu_y+\mu_x\cdot\mu_y)$

$=\sum_{i=1}^{n}x_iy_i-\mu_x\sum_{i=1}^{n}y_i-\mu_y\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\mu_x\cdot\mu_y=\sum_{i=1}^{n}x_iy_i-n\mu_x\mu_y-n\mu_x\mu_y+n\mu_x\mu_y$

$=\sum_{i=1}^{n}x_iy_i-n\mu_x\mu_y$。

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

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

$=na^2+2ab\sum_{i=1}^{n}x_i+b^2\sum_{i=1}^{n}x_i^{\,2}-2a\sum_{i=1}^{n}y_i-2b\sum_{i=1}^{n}x_iy_i+\sum_{i=1}^{n}y_i^{\,2}$

$=na^2+2abn\mu_x+b^2(S_{XX}+n\mu_x^{\,2})-2an\mu_y-2b(S_{XY}+n\mu_x\,\mu_y)+S_{YY}+n\mu_y^{\,2}$

$=n(a^2+2ab\mu_x+b^2\mu_x^{\,2})+b^2S_{XX}-2an\mu_y-2b(S_{XY}+n\mu_x\,\bar{y})+S_{YY}+n\mu_y^{\,2}$

$=n(a+b\mu_x)^2+b^2S_{XX}-2n\mu_y(a+b\mu_x)-2bS_{XY}+S_{YY}+n\mu_y^{\,2}$

$=n[(a+b\mu_x)^2-2\mu_y(a+b\mu_x)+\mu_y^{\,2}]+S_{XX}[b^2-\frac{2S_{XY}}{S_{XX}}b+(\frac{S_{XY}}{S_{XX}})^2]-\frac{S_{XY}^{\,2}}{S_{XX}}+S_{YY}$

$=n(a+b\mu_x-\mu_y)^2+S_{XX}(b-\frac{S_{XY}}{S_{XX}})^2+(S_{yy}-\frac{S_{XY}^{\,2}}{S_{XX}})$

欲使 E 的值最小$\Rightarrow\begin{cases}a+b\cdot\bar{x}-\bar{y}=0\\ b-\dfrac{S_{XY}}{S_{XX}}=0\end{cases}\Rightarrow b=\frac{S_{XY}}{S_{XX}}$，$a=\mu_y-\frac{S_{XY}}{S_{XX}}\,\mu_x$。

[方法二]：

將 X、Y 兩個變數標準化化成 X'、Y'，其中 $X'=\frac{X-\mu_x}{\sigma_x}$，$Y'=\frac{Y-\mu_y}{\sigma_y}$

X'、Y'的平均數與標準差分別爲 0 與 1

標準化後，設最佳直線 L'：$y'=a+bx'$

誤差的平方和 $E'=\sum_{i=1}^{n}[y'_i-(a+bx'_i)]^2$

$$=\sum_{i=1}^{n}[(y'_i)^2-2y'_i(a+bx'_i)+(a+bx'_i)^2]$$

$$=\sum_{i=1}^{n}(y'_i)^2-2\sum_{i=1}^{n}(ay'_i+bx'_iy'_i)+\sum_{i=1}^{n}(a^2+2abx'_i+b^2(x'_i)^2)$$

$$=\sum_{i=1}^{n}(y'_i)^2-2a\sum_{i=1}^{n}y'_i-2b\sum_{i=1}^{n}x'_iy'_i+na^2+2ab\sum_{i=1}^{n}x'_i+b^2\sum_{i=1}^{n}(x'_i)^2$$

因爲 X'、Y'的平均數與標準差分別爲 0 與 1

所以$\sum_{i=1}^{n}y'_i=\sum_{i=1}^{n}x'_i=0$

$$=\sum_{i=1}^{n}(y'_i)^2-2b\sum_{i=1}^{n}x'_iy'_i+na^2+b^2\sum_{i=1}^{n}(x'_i)^2$$

$$=\sum_{i=1}^{n}(y'_i)^2+na^2+\sum_{i=1}^{n}(x'_i)^2\ [b-\frac{\sum_{i=1}^{n}x'_iy'_i}{\sum_{i=1}^{n}(x'_i)^2}]^2-\frac{(\sum_{i=1}^{n}x'_iy'_i)^2}{\sum_{i=1}^{n}(x'_i)^2}$$

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

另一方面，$b=\frac{\sum_{i=1}^{n}x'_iy'_i}{\sum_{i=1}^{n}(x'_i)^2}=\frac{\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{S_X})(\frac{y_i-\bar{y}}{S_Y})}{\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{S_X})^2}=\frac{S_X^2}{S_XS_Y}(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$

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

$\Rightarrow$最佳直線 $y'=rx'$。

再將 $x'=\frac{x-\mu_x}{\sigma_x}$，$y'=\frac{y-\mu_y}{\sigma_y}$

代入上式

$\Rightarrow\frac{y-\mu_y}{\sigma_y}=r(\frac{x-\mu_x}{\sigma_x})\Rightarrow y-\mu_y=\frac{rS_Y}{S_X}(x-\mu_x)$

y 對 x 的最佳直線 L：$y=a+bx$，其中 $b=\frac{rS_Y}{S_X}=\frac{S_{XY}}{S_{XX}}$，$a=\mu_y-\frac{S_{XY}}{S_{XX}}\ \mu_x$

由上式可知 y 對 x 的最佳直線必過點(μ_x, μ_y)。

結論：

(1)給定 X、Y 兩個變數，如表所示 $\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ \hline Y & y_1 & y_2 & \cdots & y_n \end{array}$，將 X、Y 兩個變數標準化，化成 X'、Y'，Y'對 X'的最佳直線 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 $\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ \hline Y & y_1 & y_2 & \cdots & y_n \end{array}$，

則 Y 對 X 的最佳直線 L：$y=a+bx$ 必通過點(μ_x, μ_y)，

其中 $b=\frac{rS_Y}{S_X}=\frac{S_{XY}}{S_{XX}}$，$a=\mu_y-\frac{S_{XY}}{S_{XX}}\mu_x$。

(3)利用 Excel 求最佳直線：

指令：LINEST

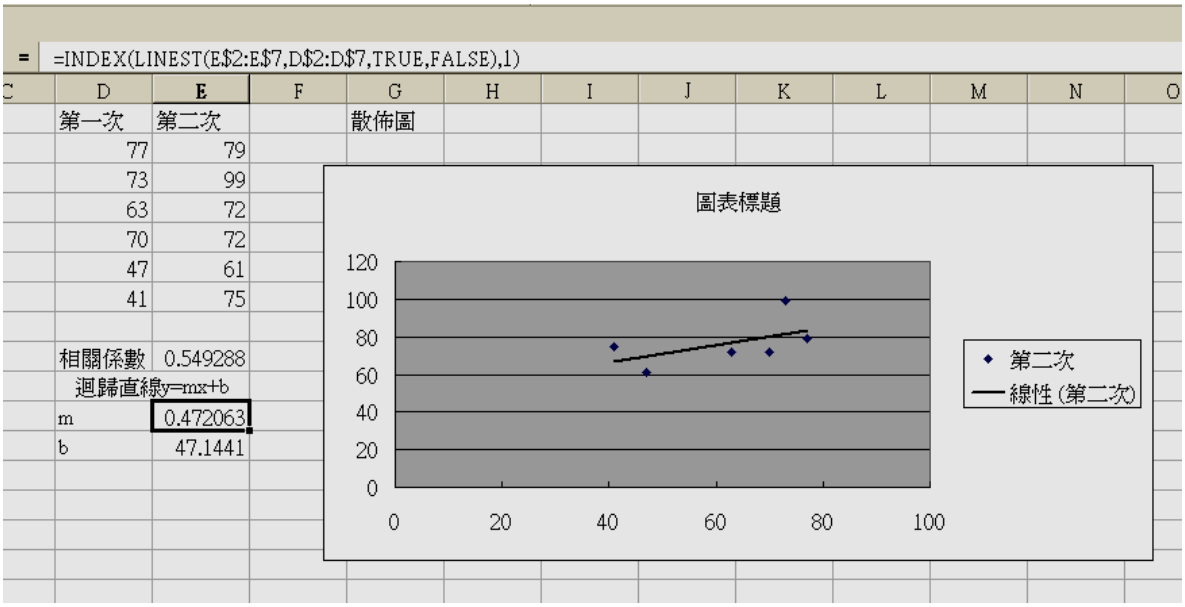
功用：使用最小平方法計算最適合於觀測資料組的迴歸直線公式，並傳回該直線公式的陣列。由於此函數傳回陣列值，所以必須輸入爲陣列公式。

語法：　LINEST(known_y's,known_x's,const,stats)

最佳直線：$y=mx+b$

m 的計算：INDEX(**LINEST(known_y's**,known_x's,const,stats),1)

b 的計算：INDEX(**LINEST(known_y's**,known_x's,const,stats),2)



[**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。

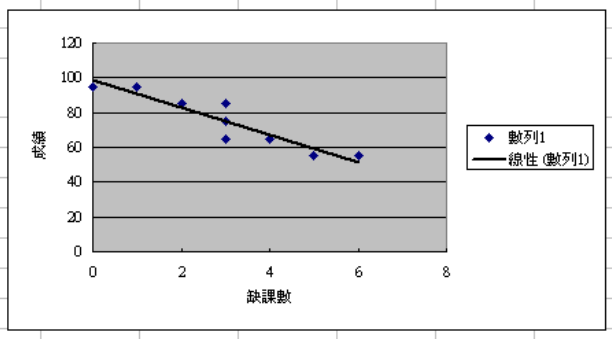
(2)求這 10 個資料變數 Y 對變數 X 的最佳直線方程式。

(3)根據這條最佳直線，請預測缺課數爲 7 時的成績爲多少？

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

B7　=　=INDEX(LINEST(B$3:K$3,B$2:K$2,TRUE,FALSE),1)

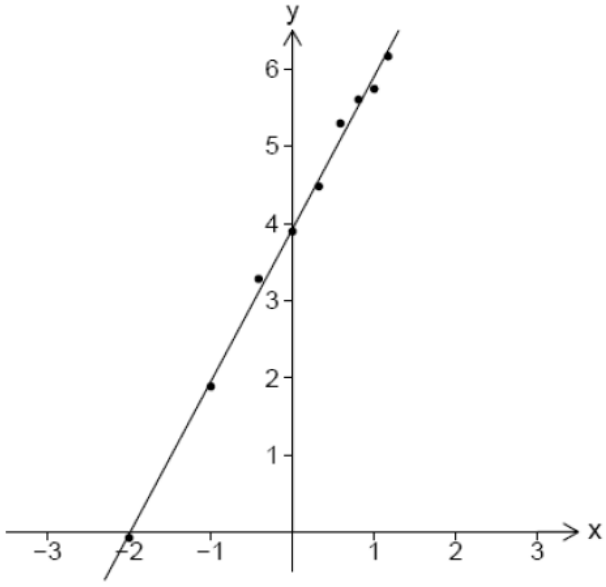
	A	B	C	D	E	F	G	H	I	J	K
1	編號	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	缺課數	0	3	6	5	3	4	3	3	2	1
3	成績	95	75	55	55	65	65	75	85	85	95
4											
5	相關係數	-0.92967									
6	迴歸直線 y=mx+b										
7	m	-7.85714									
8	b	98.57143									
9											
10											
11											
12											
13											
14											



[例題7] 某人進行一實驗來確定某運動之距離 d 與時間 t 的平方或立方成正比，所得數據如下：

時間 t(秒)	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25
距離 d(呎)	0.95	3.69	9.71	14.88	22.32	39.34	48.68	53.65	71.79

爲探索該運動的距離與時間之關係，令 $x=\log_2 t$，$y=\log_2 d$，即將上述數據(t,d)分別取以 2 爲底的對數變換，例如：(2 , 53.65)變換成爲(2 , 5.74)。已知變換後的數據$(x_1,y_1),(x_2,y_2),\ldots,(x_9,y_9)$之散布圖及以最小平方法所求得變數 y 對變數 x 的最適合直線(或稱迴歸直線)爲 $y=a+bx$，如下圖所示：



試問下列哪些選項是正確的？
(1)若 $d=14.88$，則 $3<\log_2 d<4$
(2)x 與 y 的相關係數小於 0.2
(3)由上圖可以觀察出 $b>2.5$
(4)由上圖可以觀察出 $a>2$
(5)由上圖可以確定此運動之距離與時間的立方約略成正比。
Ans：(1)(4)　(2008 指定甲)

(練習8) 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
$\overline{x}=65, \overline{y}=70, S_X=10, S_Y=5, r=0.8$
(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9) 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10) 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

(練習11) 請利用 EXCEL 完成問題：
1976 年 Marc 和 Helen Bornstein 研究了日常生活的步調，觀察城鎮的規模變大之後，生活節奏是否變快。他們有系統地觀察了城鎮主要街道上徒步者步行 50 英尺所須的平均時間，下表是他們蒐集的數據，V 代表步行 50 英尺的平均速率，P 代表城鎮人口，

地區	步行平均速率 V(英尺/秒)	人口(P)
布爾諾(捷克)	4.81	341948
布拉格(捷克)	5.88	1092759
科特(科西嘉)	3.31	5491
巴士底(法國)	4.90	49375
慕尼黑(德國)	5.62	1340000
塞克農克里特(希臘)	2.76	365
依提雅(希臘)	2.27	2500
伊拉克林(希臘)	3.85	78200
雅典(希臘)	5.21	867023
沙非特(以色列)	3.70	14000
戴姆拉(以色列)	3.27	23700
納塔尼亞(以色列)	4.31	70700
耶路撒冷(以色列)	4.42	304500
新海文(美國)	4.39	138000
布魯克林(美國)	5.05	602000

(1)請建立一個 logV 對 logP 的數據表。

(2)利用 Excel 作出(1)中數據表的散布圖。

(3)若 P 與 V 的關係可以用 $P=CV^{\alpha}$來表示，請估計 C 與α的值約爲多少。

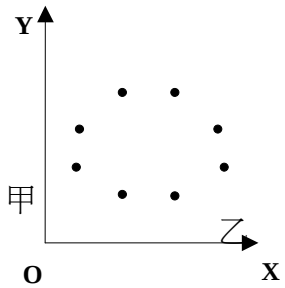
綜合練習

(1) 某校高三共有 300 位學生， 數學科第一次段考、第二次段考成績分別以 X、Y 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爲 0.016， 試問下列哪些選項是正確的？
(1) X 與 Y 的相關情形可以用散佈圖表示
(2)這兩次段考的數學成績適合用直線 $y=a+bx$ 表示 X 與 Y 的相關情形 (a,b 爲常數，$b\neq 0$)
(3)X+5 與 Y+5 的相關係數仍爲 0.016。
(4)10X 與 10Y 的相關係數仍爲 0.016。
(5)若 $X'=\frac{X-\overline{X}}{S_X}$ ，$Y'=\frac{Y-\overline{Y}}{S_Y}$ ，其中 $\overline{X}$ 、$\overline{Y}$ 分別爲 X、Y 的平均數，S_X、S_Y 分別爲 X、Y 的標準差，則 X'與 Y'的相關係數仍爲 0.016。
(2007 指定甲)

(2) 經濟學者分析某公司服務年資相近的員工之「年薪」與「就學年數」的資料，得到這樣的結論：『員工就學年數每增加一年，其年薪平均增加 8 萬 5 千元』。試問上述結論可直接從下列哪些選項中的統計量得到？
(1) 「年薪」之眾數與「就學年數」之眾數
(2) 「年薪」之全距與「就學年數」之全距
(3) 「年薪」之平均數與「就學年數」之平均數
(4) 「年薪」與「就學年數」之相關係數
(5) 「年薪」對「就學年數」之迴歸直線斜率 (2009 指定乙)

(3) 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)變異數 (E)全距
(88 自)

(4) 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)



(5) 某班的 50 名學生參加一項考試，考題共 100 題，全爲 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爲 1。 (90 自)

(6) 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

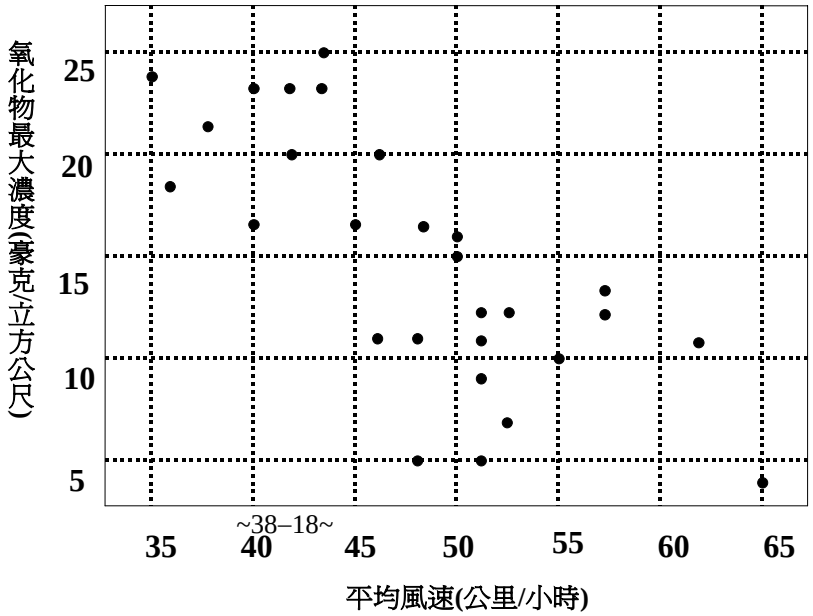
(7) 數學老師想把某次模擬考滿分 120 分的成績(X)作調整爲滿分 100 分的平時成績(Y)，以便登記成一次平時成績，故 $Y=\frac{5}{6}\cdot X$。現在模擬考的成績求得算術平均數 $\bar{x}$，中位數 M*e*，全距 D，標準差 S，數學與物理分數相關係數爲 *r*；若調整之後，各相對統計量爲算術平均數 $\bar{x}'$，中位數 M*e*′，全距 D′，標準差 S′，數學與物理分數相關係數爲 *r*′，則下列何者正確？(A) $\bar{x}'=\frac{5}{6}x$ (B)M*e*=M*e*′ (C)$D'=\frac{5}{6}$D (D)$S'=\frac{5}{6}$S (E)*r*=*r*′。

(8) 十位考生之國文與數學成績列表如下：

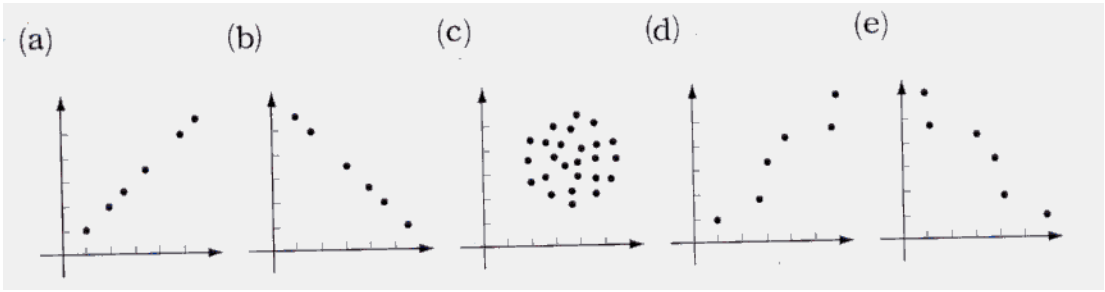
考生編號	1	2	3	4	5	6	7	8	9	10
國文	89	65	76	69	82	57	66	72	78	66
數學	75	57	65	65	83	63	58	62	63	69

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

(9) 空氣品質會受到污染物排放量及大氣擴散等因素的影響。某一機構爲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爲 15。
(C)此筆資料，平均風速的中位數介於 45 與 50 之間。
(D)若以最小平方法決定數據集中直線趨勢的直線，則該直線的斜率小於 0。
(2002 指定甲)



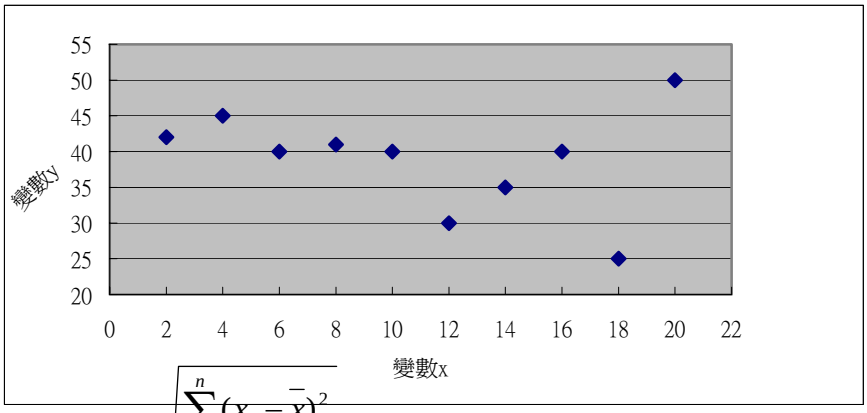
(10) 請排出下面 5 個散布圖中 x,y 的相關係數的大小順序。



(11) 如下表，爲 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 位同學的平均成績爲 12 分，變異係數爲 $\frac{50\sqrt{3}}{9}$ %，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) 右圖爲兩變數(x,y)10 筆資料的散布圖，
試問下列敘述那些是正確的？



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

(A)變數 y 的中位數爲 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 爲正相關
(E)變數 x,y 的最佳直線斜率爲負的

(13) 某一個樣本共有 100 筆資料，經計算已知如下資料：

$\sum_{i=1}^{100}x_i=12500$，$\sum_{i=1}^{100}y_i=8000$，$\sum_{i=1}^{100}x_i^2=1585000$，$\sum_{i=1}^{100}y_i^2=648100$，$\sum_{i=1}^{100}x_iy_i=1007423$

(a) $\bar{x}=?$　　$\bar{y}=?$

(b)S_X^2=？　S_Y^2=？
(c)相關係數 r=？
(d)求 y 對 x 的最佳直線方程式。
(e)若將每個 x_i 乘以 2 再加上 50，每個 y_i 乘以 2 再加上 100，可得新的數據 x_i'、y_i'，求新的數據 x_i'、y_i'的相關係數。

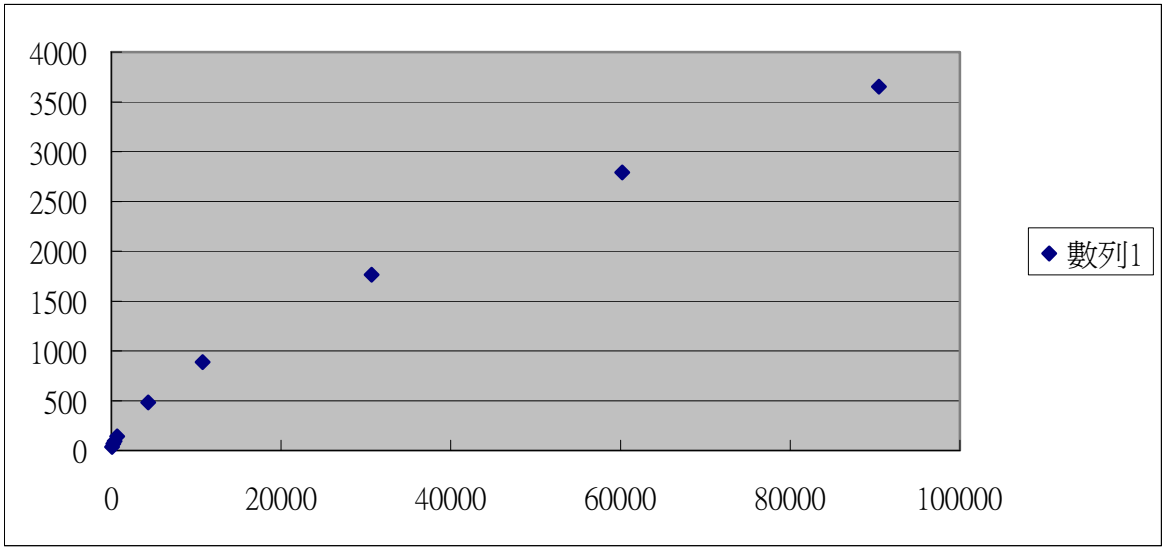
(14) 下表是太陽系九大行星的週期與到太陽的平均距離：
我們想建立一個數學模型來描述週期與平均距離間的關係。

行星	週期(天)	平均距離(百萬英里)
水星	88.0	36
金星	224.7	67.25
地球	365.3	93
火星	687.0	141.75
木星	4331.8	483.80
土星	10760.0	887.97
天王星	30684.0	1764.50
海王星	60188.3	2791.05
冥王星	90466.8	3653.90

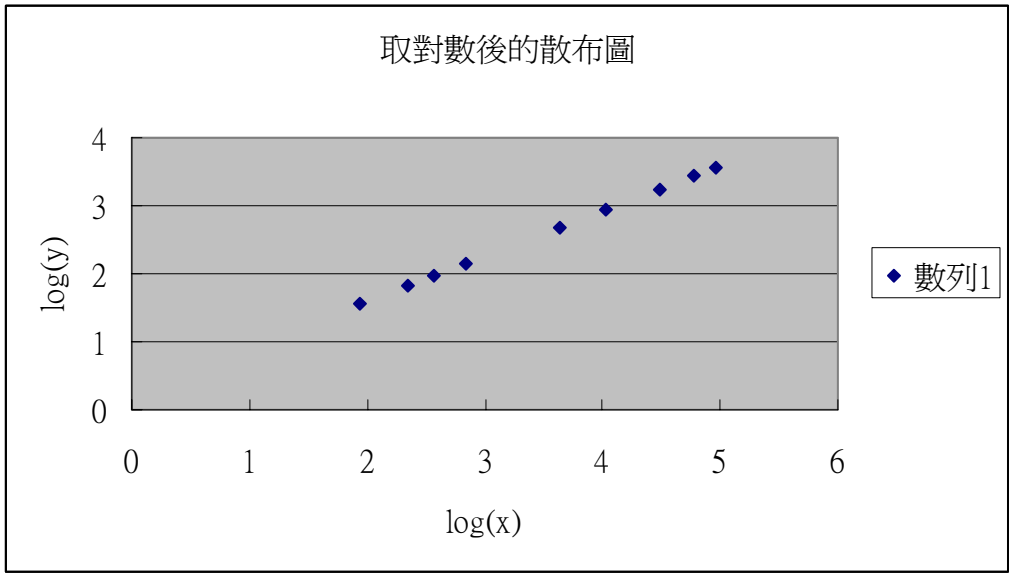
(a)用 EXCEL 可以畫出平均距離對週期的散布圖
(b)將周期(x)與平均距離(y)取對數，令 Y=log$y$，X=log$x$，用 EXCELX 畫出 X 對 Y 的散布圖
(c)利用 EXCEL 計算 X、Y 的相關係數與最佳直線。
(d)利用上述的資料，請找出周期(x)與平均距離(y)的關係。

綜合練習解答

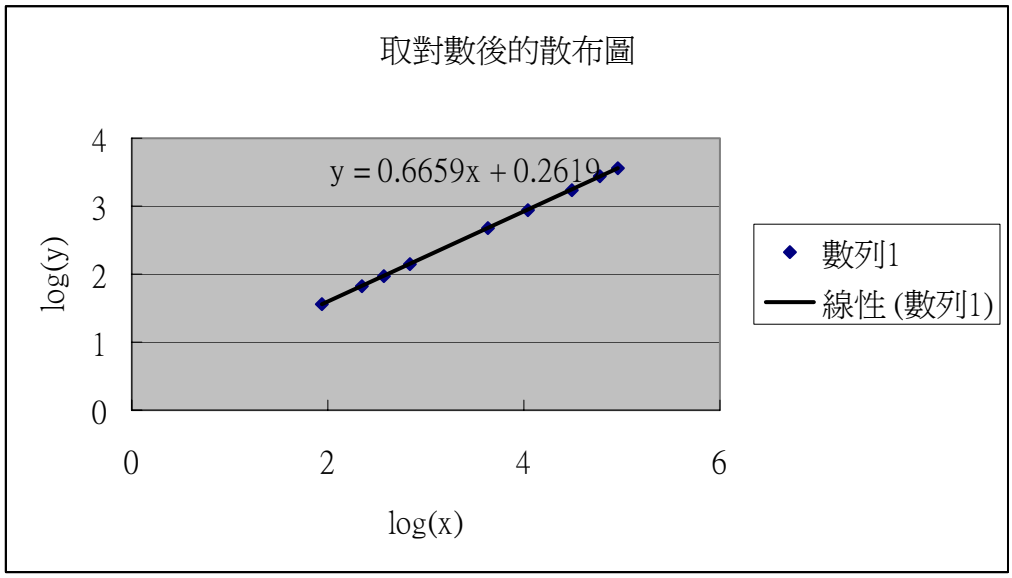
(1) (1)(3)(4)(5)
(2) (5)
(3) (C)(E)
(4) (C)
(5) (A)(B)(D)(E)
(6) (B)
(7) (A)(D)(E)
(8) (C)
(9) (C)(D)
(10) (a)>(d)>(c)>(e)>(b)
(11) $x=14$，$y=10$
(12) (A)(B)(C)(E)
(13) (a)125，80 (b)227.27，81.82 (c)0.55 (d)$y=0.33x+38.76$ (e)0.55
(14) (a)



(b)



(c)



Y=0.6659X+0.2619，相關係數=0.99998，(d) $y=a\cdot x^{0.6619}\approx a\cdot x^{\frac{2}{3}}$