重點 1：散布圖(scatter diagram)

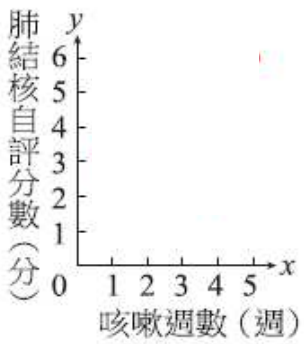
1.意義：探討兩個變量：如(身體，體重)、(容量，體積)、(風力，雨量)、(數學成績，國文成績)、(BMI 指數，血壓)等之間是否有關聯，稱為**二維數據分析**。

2.定義：兩個可能相關的變量數據(X，Y)，將第一個變量當作 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，所得的圖形稱為此兩個變量數據的**散佈圖**。

註：數據 X 表示$(x_1, x_2, \cdots, x_n)$，Y 表示$(y_1, y_2, \cdots, y_n)$

例 1.1：醫生統計六位病患的咳嗽週數 x (週)與肺結核自評分數 y (分)如下表。繪出此數據的散布圖。

民眾	甲	乙	丙	丁	戊	己
x (週)	2	1	4	5	4	2
y (分)	3	3	5	6	2	5



重點 2：直線相關概述

1.意義：檢視散布圖中，可發現數據點(x_i, y_i)分布集中在某一條直線 L 的附近，稱點(x, y)為直線相關

註：若點的分布呈現曲線形狀時，表示兩變數之間的關係非線性關係，不在高中討論的範圍內。

2.直線相關性質：

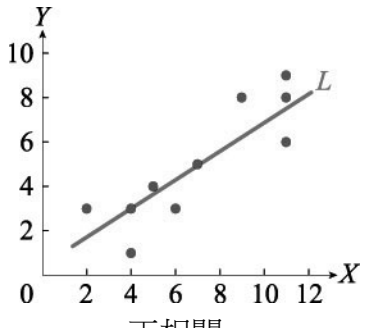
(1)正相關：當直線 L 的斜率為正時，稱變數 X 與 Y 為正相關，即兩個變量有**一致的趨勢**(同時增加或減少)

(2)完全正相關：數據點(x_i, y_i)全部在一條**斜率為正**的直線上

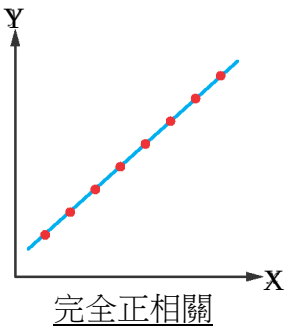
(3)負相關：當直線 L 的斜率為負時，稱變數 X 與 Y 為負相關，即兩個變量**趨勢相反**，一個增加(減少)，則另一個就減少(增加)

(4)完全負相關：數據點(x_i, y_i)全部在一條**斜率為負**的直線上

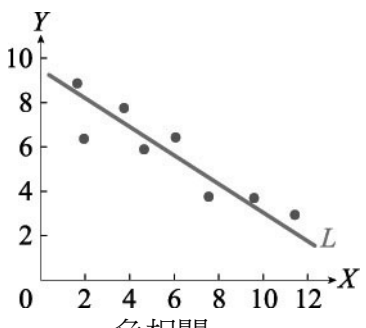
(5)零相關：一個變量的變化對另一個變量沒有影響。當散布的各點，**上下左右均成對稱狀態**，或各點完全分布在**平行 x 軸**或**平行 y 軸**的直線上，稱兩變量 X 與 Y 為零相關



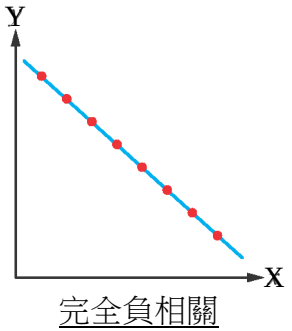
正相關



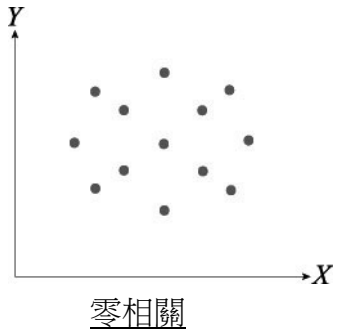
完全正相關



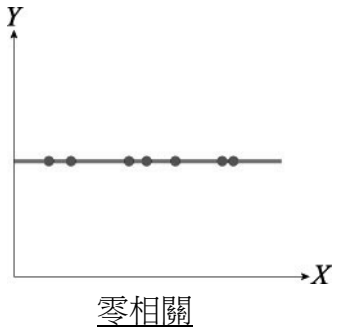
負相關



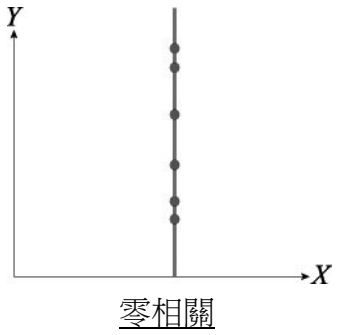
完全負相關



零相關



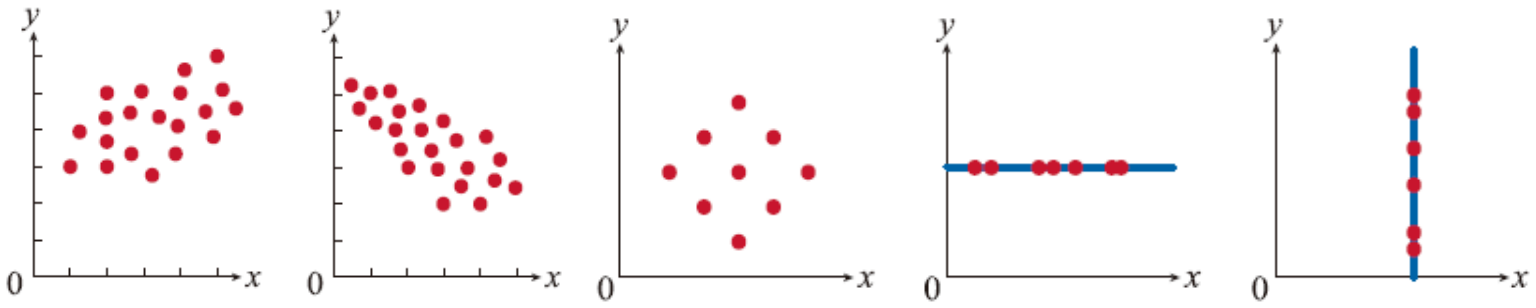
零相關



零相關

例 2.1：觀察下列各圖中兩變量 x 與 y 之間的相關情形，試說明其相關性質：

(A)　(B)　(C)　(D)　(E)



重點 3：數據標準化(standardized data)

1.意義：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。因此會先將數據標準化後，再描繪出其散佈圖，以消彌不同單位與刻度之散佈圖的差異性。

2.二維數據標準化定義：

設數據 X：x_1，x_2，…，x_n，平均數為 μ_x，標準化 $x_i' = \dfrac{x_i - \mu_x}{\sigma_x}$

數據 Y：y_1，y_2，…，y_n，平均數為 μ_y，標準化 $y_i' = \dfrac{y_i - \mu_y}{\sigma_y}$

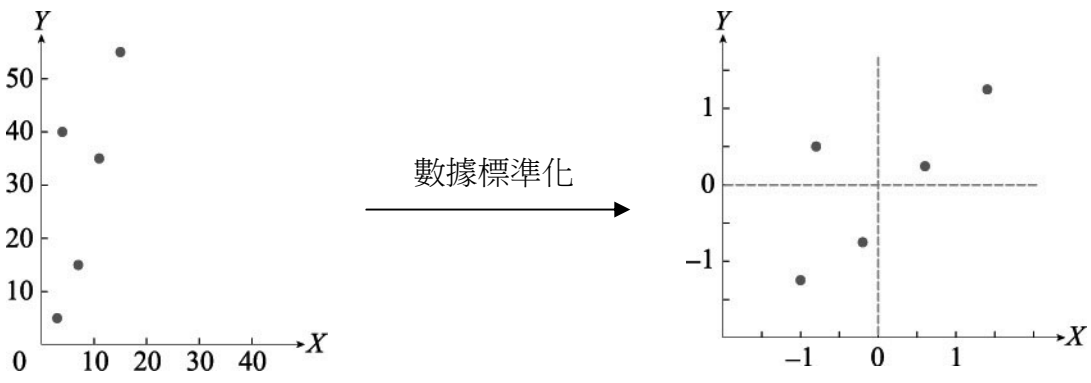
再將已標準化後的各點坐標 (x_1', y_1')，(x_2', y_2')，…，(x_n', y_n') 描繪在坐標平面上，所得的圖形稱為數據標準化的**散佈圖**。

3.數據標準化的性質：

(1)數據標準化後的變量都**沒有單位**

(2)標準化後，意即將數據轉換成平均數為 0，標準差為 1 的新數據，以減少不同測量單位的數據對圖形的影響，並使數據的分布情形更加明顯

(3)原來兩變量的**平均值**經過標準化後變成**原點**，則兩變量的圖形就可以重疊在一起比較



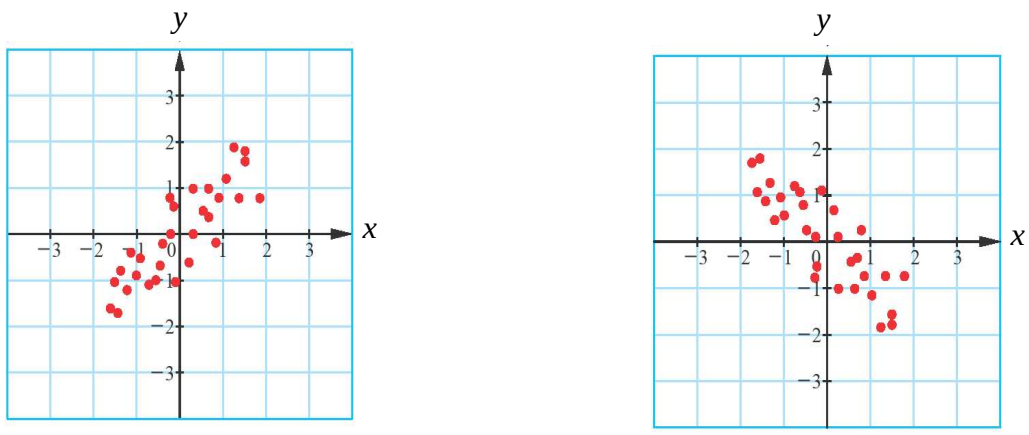
原始數據散布圖　　標準化數據散布圖

4.數據標準化散的佈圖特性：

(1)設已標準化後的各點坐標為 (x_1', y_1')，(x_2', y_2')，…，(x_n', y_n')，則其散布圖為在坐標上加畫兩條**平均線**，亦即以平均數為原點，將全圖分為四個象限。

(2)若數據點(x_i，y_i)在第一與第三象限內時，則各點的 $x_i' y_i'$ **乘積為正**。

若數據點(x_i，y_i)在第二與第四象限內時，則各點的 $x_i' y_i'$ **乘積為負**。



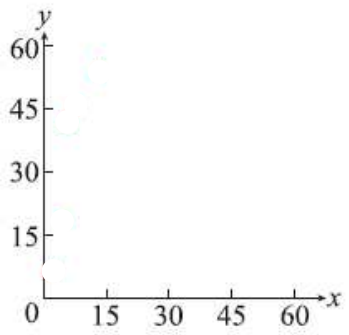
正相關

負相關

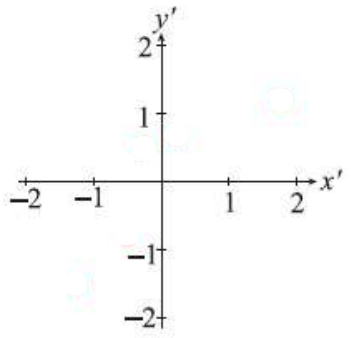
例 3.1：已知兩變量 x 與 y 的 5 筆數據如下表，試將數據標準化，並繪出標準化數據的散佈圖：

x	1	4	5	7	13
y	6	18	42	45	54

$x'=\dfrac{x_i-\mu_x}{\sigma_x}$					
$y'=\dfrac{y_i-\mu_y}{\sigma_y}$					



原始數據



標準化數據

重點 4：相關係數(correlation coefficient)

1.意義：從散布圖中可呈現兩個變量間關聯的方向、形式和強度等之相關性，而精確的描述兩變量的相關程度的高低，稱做**相關係數**，一般以 r 表示。

註：相關係數是採用皮爾森積差相關係數(Pearson Product-Moment Correlation Coefficient)

2.定義：

(1)數據標準化：

設數據 X：x_1，x_2，…，x_n，平均數為 μ_x，標準差為 σ_x，標準化數據為 $x_i'=\dfrac{x_i-\mu_x}{\sigma_x}$

數據 Y：y_1，y_2，…，y_n，平均數為 μ_y，標準差為 σ_y，標準化數據為 $y_i'=\dfrac{y_i-\mu_y}{\sigma_y}$

則數據標準化數據(x_i'，y_i')之**相關係數** $r=\dfrac{1}{n}\sum_{i=1}^{n}x_i'y_i'$ 背 (標準化公式)

(2)數據未標準化：

數據 X 的標準差 $\sigma_x=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{n}}=\sqrt{\dfrac{S_{xx}}{n}}$　數據 Y 的標準差 $\sigma_y=\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\mu_y)^2}{n}}=\sqrt{\dfrac{S_{yy}}{n}}$

則：**相關係數** $r=\dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{x_i-\mu_x}{\sigma_x}\right)\left(\dfrac{y_i-\mu_y}{\sigma_y}\right)=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n\cdot\sigma_x\cdot\sigma_y}$ 背 ….(未標準化公式 1)

或**相關係數** $r=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\cdot\sum_{i=1}^{n}(y_i-\mu_y)^2}}=\dfrac{S_{xy}}{\sqrt{S_{xx}\sqrt{S_{yy}}}}=\dfrac{\sum_{i=1}^{n}x_iy_i-n\mu_x\mu_y}{\sqrt{\sum_{i=1}^{n}(x_i^2-\mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i^2-\mu_y)^2}}$ 背 ….(未標準化公式 2)

◎標準化數據的相關係數

例 4.0：設一數據的標準化數據如下表，試求其相關係數。(取到小數點後第四位)

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

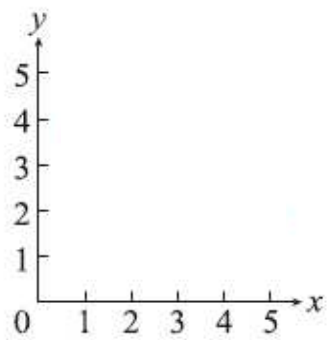
◎未標準化數據的相關係數

例 4.1：兩變量 x 與 y 的數據如下表：

(1)繪製 x 與 y 的散布圖

(2)求 x 與 y 的相關係數

x	1	2	3	4	5
y	4	5	3	1	2

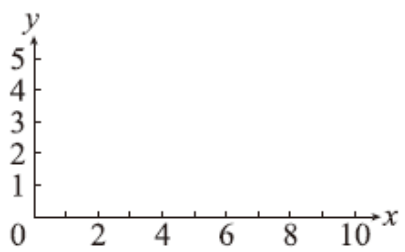


例 4.2：兩變量 x 與 y 的數據如下表：

(1)繪製 x 與 y 的散布圖

(2)求 x 與 y 的相關係數

x	2	4	6	8	10
y	1	2	3	4	5



重點 5：相關係數的性質

1.相關係數 r 的範圍：$-1 \le r \le 1$

(1)完全正相關：相關係數 $r=1$，即各點恰落在一條斜率為正的直線上

(2)正相關：相關係數 r，$0<r<1$，細分如下：

高度正相關 $0.7 \le r<1$　　中度正相關 $0.3 \le r<0.7$　　低度正相關 $0<r<0.3$

(3)零相關：相關係數 $r=0$

(4)負相關：相關係數 r，$-1<r<0$，細分如下：

高度負相關 $-1<r \le -0.7$　　中度負相關 $-0.7<r \le -0.3$　　低度負相關 $-0.3<r<0$

(5)完全負相關：相關係數 $r=-1$，即各點恰落在一條斜率為負的直線上

2.有一組二維數據 (x_1, y_1)，(x_2, y_1)，…，(x_n, y_n)，滿足 $y_i = a x_i + b$，其中 a，b 為常數，則：

(1)若 $a>0$，則此二維數據相關係數 $r=1$

(2)若 $a<0$，則此二維數據相關係數 $r=-1$

3.有一組二維數據(x_1, y_1)，(x_2, y_1)，…，(x_n, y_n)，其相關係數為 r，

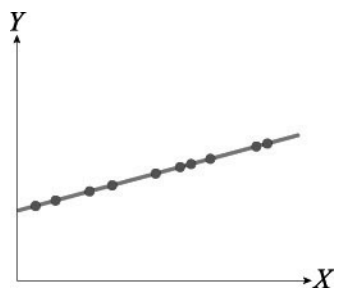
另一組二維數據(x_1', y_1')，(x_2', y_1')，…，(x_n', y_n')，滿足 $x_i'=a x_i+b$，$y_i'=c y_i+d$，其中 a，b，c，d 為常數

設新數據的相關係數為 r'，則：

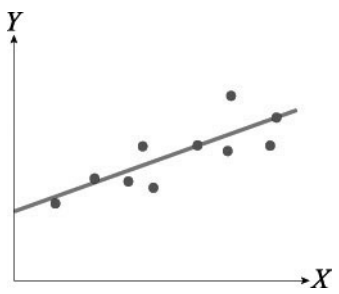
(1)若 $ac>0$，則此二維數據相關係數 $r'=r$

(2)若 $ac<0$，則此二維數據相關係數 $r'=-r$

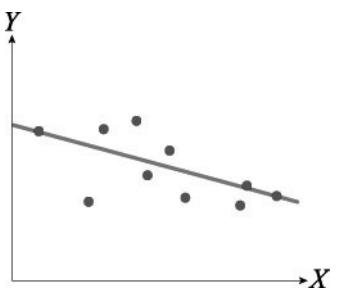
註：$|r|$愈大表示兩變量間的相關程度愈強



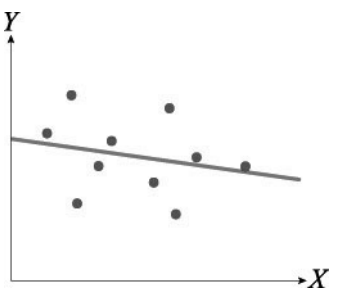
完全正相關



高度正相關



中度負相關



低度負相關

4.性質：

(1)相關係數 r 與單位無關，即改變 X 和 Y 的度量單位時，它們的相關係數並不會改變

(2)數據 X 和 Y 互相對調，其相關係數不會改變

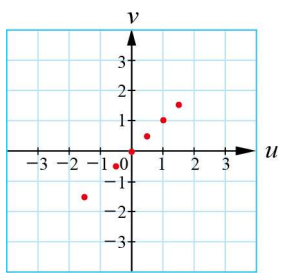
(3)相關係數、平均數、標準差等皆容易受少數極端數據值影響

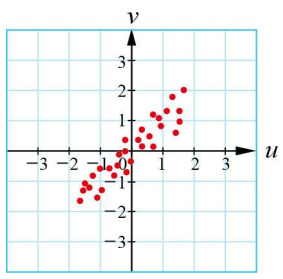
(4)相關係數只顯現兩變數之間的線性關聯性強弱，**不能作成兩變數之間的推論**，即使兩個變量的相關係數很高，這兩個變量也不一定有因果關係，需要對整體(如數據)狀況有進一步的了解之後，才能下定論

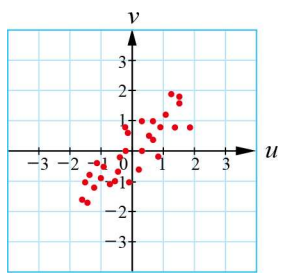
例如：「咖啡因攝取量與心臟病罹患率」的相關係數很高(高度正相關)，卻不能作成「喝咖啡容易得心臟病」的推論

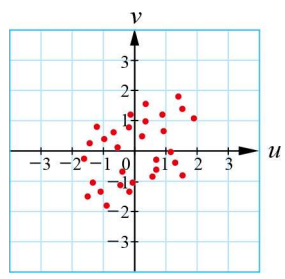
註：散佈圖(數據標準化與否)和相關係數的關係示例：

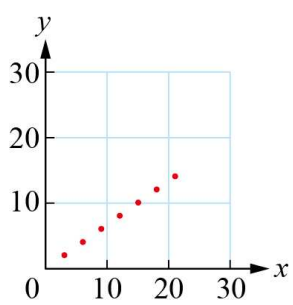
(1)相關**係數 $r>0$**



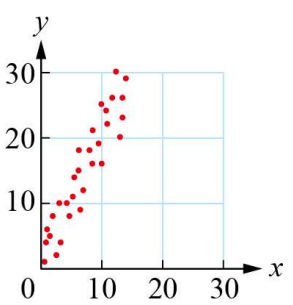




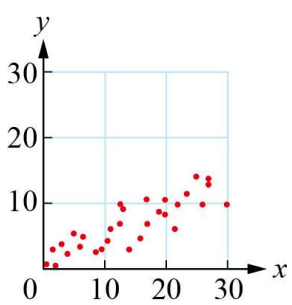




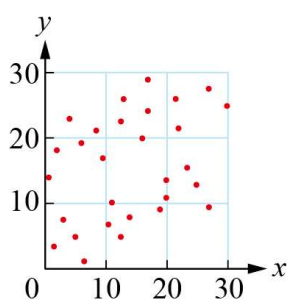
$r=1$



$r=0.9$

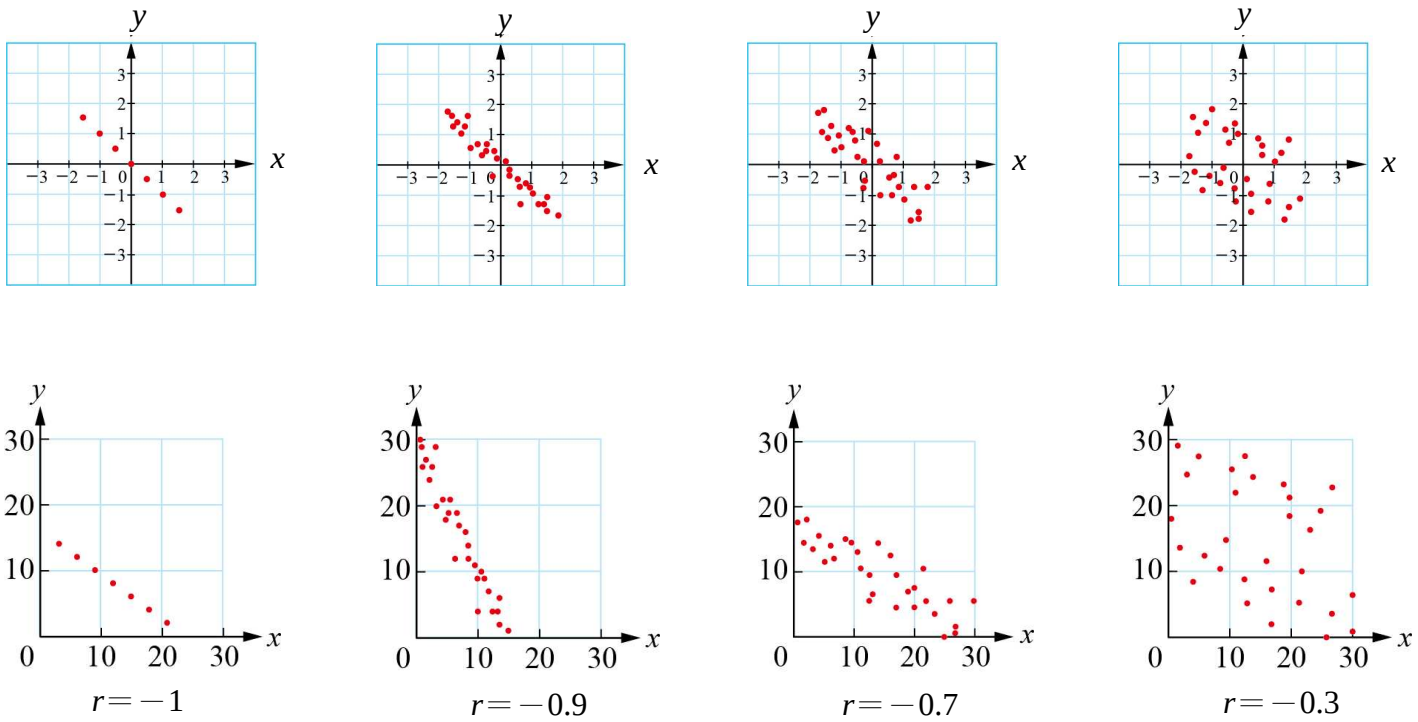


$r=0.7$



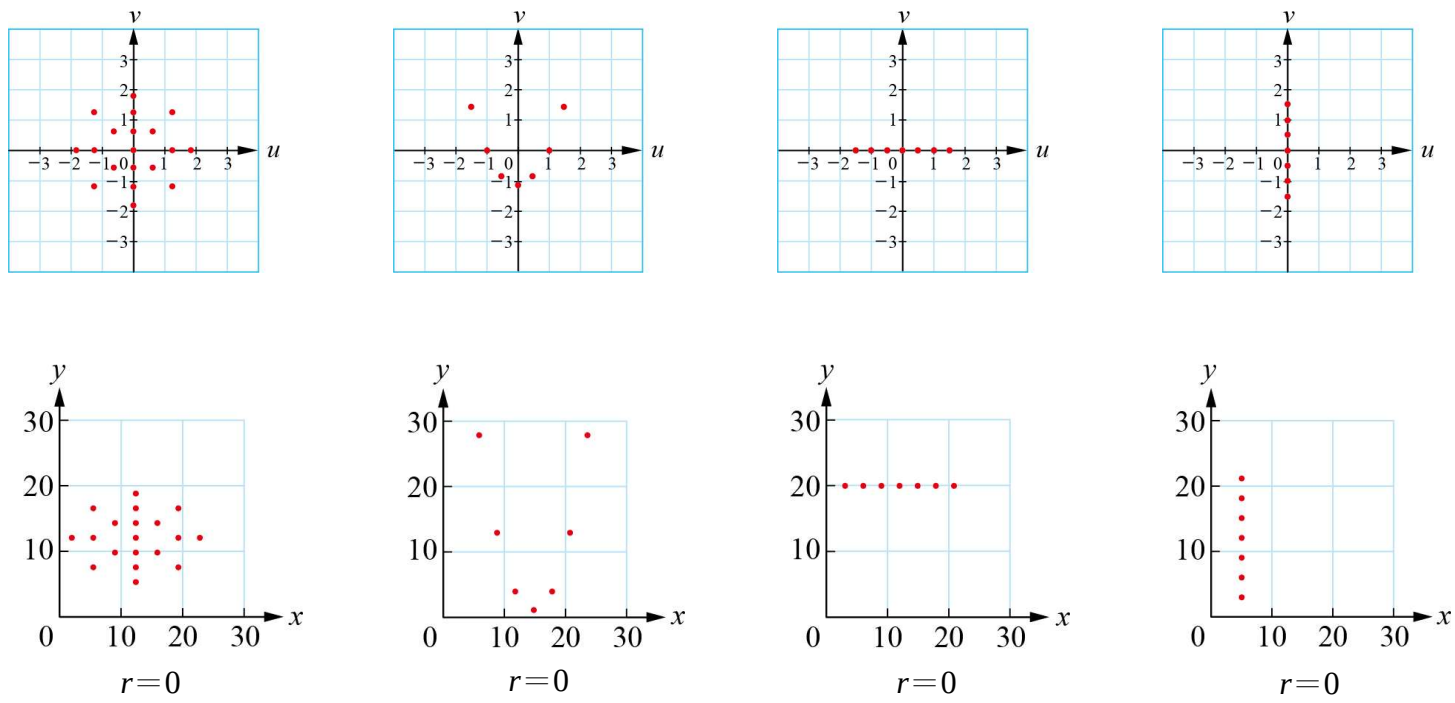
$r=0.3$

(2)相關係數 $r<0$



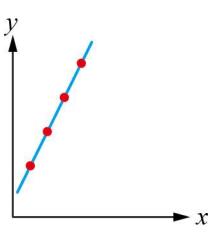
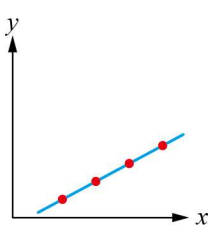
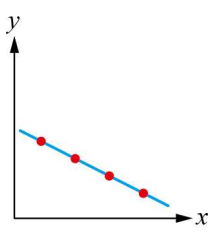
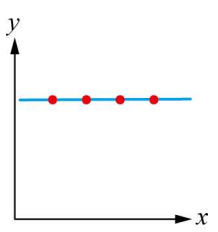
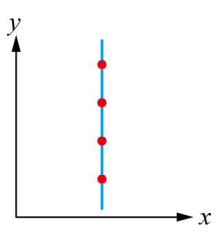
$r=-1$　　$r=-0.9$　　$r=-0.7$　　$r=-0.3$

(3)相關係數 $r=0$



$r=0$　　$r=0$　　$r=0$　　$r=0$

例 5.1：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) 

重點 6：最小平方法與迴歸直線(或稱最適直線、最佳直線)

1.意義：當散布圖顯示出兩變量 X 與 Y 間為直線相關時，希望找出一條「最適合代表兩變量之間關係的直線 L」，稱為兩變量 X 與 Y 的迴歸直線，進而利用變數 X 來預測變數 Y 的值。

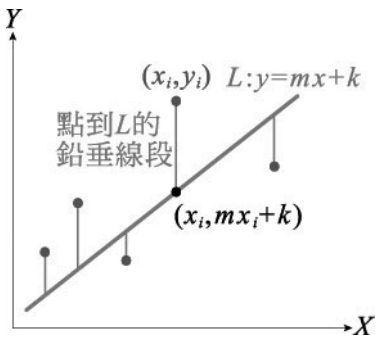
2.最小平方法定義：

A 緣由：「最適合代表兩變量之間關係的直線 L」的求法很多種，考慮計算上的方便，採用由數學家高斯所提出的「**最小平方法**」。

B 定義：

(1)設迴歸直線 $L:y=mx+k$ ，將數據 $(x_i,\ y_i)$ 與點 (x_i, mx_i+k) 的距離 $\left|y_i-(mx_i+k)\right|$ ，如右圖

稱為點 $(x_i,\ y_i)$ 到直線 L 的**鉛垂線段長度**



(2)當有 n 筆數據 $(x_1,\ y_1)$ ，$(x_2,\ y_2)$ ，…，$(x_n,\ y_n)$ 時，會有 n 筆鉛垂線段長度

$\left|y_1-(mx_1+k)\right|$ ，$\left|y_2-(mx_2+k)\right|$ ，…，$\left|y_n-(mx_n+k)\right|$

(3)計算此 n 個鉛垂線段長度的**平方和**之最小值，稱為**最小平方法**

即 $Q=\left[y_1-(mx_1+k)\right]^2+\left[y_2-(mx_2+k)\right]^2+\cdots+\left[y_n-(mx_n+k)\right]^2$

註：意即求數據點到直線 L 之鉛垂距離的平方和的最小值

3.迴歸直線的求法：

(1)若數據**標準化**為 $\left(x_i',\ y_i'\right)$ ，則迴歸直線方程式為 $y'=rx'$ ，$x'=\dfrac{x-\mu_x}{\sigma_x}$ ，$y'=\dfrac{y-\mu_y}{\sigma_y}$ ….(公式 1)　背

註：迴歸直線直線是通過原點 O(0，0)的直線，且斜率 m 恰好就是相關係數 r

(2)若數據**未標準化**，則迴歸直線方程式為 $y-\mu_y=m(x-\mu_x)$ ，斜率 $m=r\cdot\dfrac{\sigma_y}{\sigma_x}=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sum_{i=1}^{n}(x_i-\mu_x)^2}=\dfrac{S_{xy}}{S_{xx}}$ …(公式 2)　背　背

意即由(1) $y'=rx'$ ，得 $y-\mu_y=r\cdot\dfrac{\sigma_y}{\sigma_x}(x-\mu_x)=\dfrac{S_{xy}}{S_{xx}}(x-\mu_x)$　背

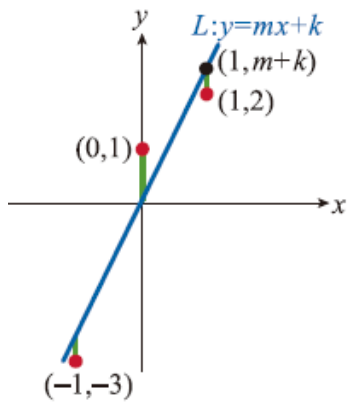
4.迴歸直線的性質：

(1)迴歸直線**必通過算術平均數**的點 $(\mu_x\ ,\ \mu_y)$

(2)相關係數和迴歸直線式是密切相關的，相關係數度量直線相關的方向和強度，而迴歸直線描述此相關

◎迴歸直線的定義

例 6.1：散佈圖上有資料(0，1)，(1，2)，(－1，－3)，試利用最小平方法求迴歸直線方程式。



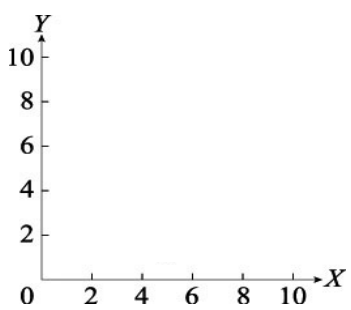
例 6.2：某系申請入學考生的筆試成績似乎與口試有關。老師隨機選了 5 位考生，其筆試與口試成績如下：

考生	甲	乙	丙	丁	戊
筆試成績 x(分)	5	5	4	7	9
口試成績 y(分)	3	1	4	3	9

(1)試作散佈圖

(2)求此五位同學 x 與 y 的相關係數

(3)求 y 對 x 的迴歸直線方程



(2)

	x	y	$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)^2$	$(y-\mu_y)^2$	$(x-\mu_x)(y-\mu_y)$
甲							
乙							
丙							
丁							
戊							

◎預測

例 6.3：飲料店調閱上個月某四天的當日最高氣溫 x (°C)與銷售金額 y (千元)如下表：

最高氣溫 x	33	31	29	27
銷售金額 y	14	12	8	10

(1)試求 y 對 x 的迴歸直線方程式

(2)利用迴歸直線預測：當最高氣溫為 35°C 時，銷售金額為多少元？

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

例 6.4：已知變量 x 的平均數 $\mu_x=6$，標準差 $\sigma_x=3$；變量 y 的平均數 $\mu_y=9$，標準差 $\sigma_y=5$，且 x 與 y 的相關係數為 -0.8，求 y 對 x 的迴歸直線方程式。

◎迴歸直線必過(μ_x，μ_y)

例 6.5：已知兩變數 X、Y 的數據如右：

若以最小平方法求得 Y 對 X 的迴歸直線為 $y=-\frac{1}{2}x+\frac{11}{2}$，求 a 之值。

X	1	2	3	6
Y	4	5	a	2

◎可以使用計算機或電腦軟體來輔助計算

例 6.6：玉山觀測站提供的 2017 年每月氣溫 x (°C)與 2018 年每月氣溫 y (°C)如下表：

月份	1	2	3	4	5	6	7	8	9	10	11	12
2017 年氣溫 x	1.6	0.1	0.3	3.4	5.4	7.6	8.5	9.2	9.7	8.3	5	2.1
2018 年氣溫 y	0.5	－0.3	1.4	4.8	8	8.3	7.6	7.2	7.1	4.8	4.4	5.1

利用電腦軟體 Excel，求：

(1)這兩年氣溫的相關係數。(四捨五入到小數點以下第 1 位)

(2) y 對 x 的迴歸直線方程式。

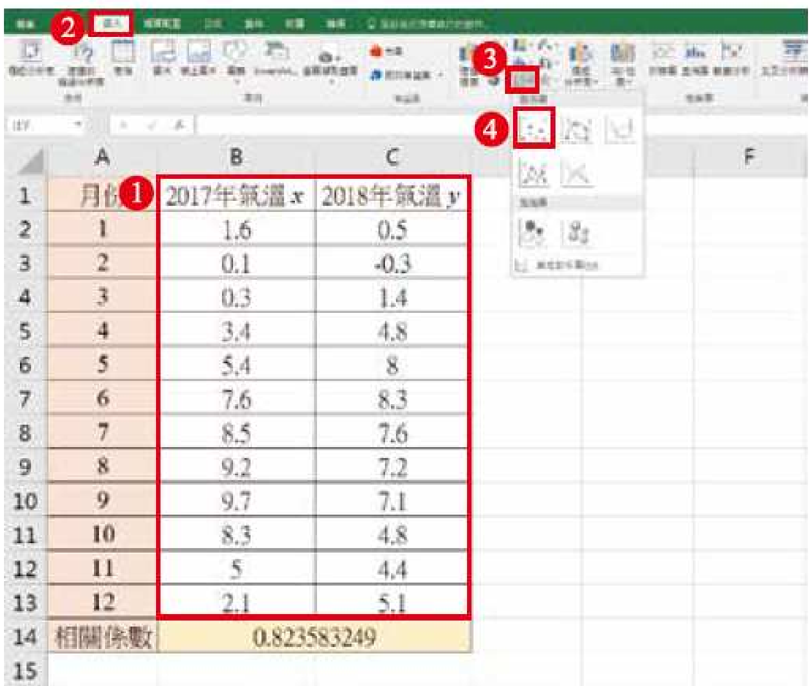
解：(1)

	A	B	C
1	月份	2017年氣溫 x	2018年氣溫 y
2	1	1.6	0.5
3	2	0.1	-0.3
4	3	0.3	1.4
5	4	3.4	4.8
6	5	5.4	8
7	6	7.6	8.3
8	7	8.5	7.6
9	8	9.2	7.2
10	9	9.7	7.1
11	10	8.3	4.8
12	11	5	4.4
13	12	2.1	5.1
14	相關係數	0.823583249	

❶將題目中的月份、2017 年氣溫與 2018 年氣溫分別輸入 Excel 的 A 欄、B 欄與 C 欄

❷在欲顯示相關係數的儲存格輸入=CORREL(B2:B13,C2:C13)，
即計算儲存格 B2:B13 和 C2:C13 的相關係數。故這兩年氣溫的相關係數為 0.8

(2)我們可以透過製作散布圖來求 y 對 x 的迴歸直線方程式，如下圖所示。



	A	B	C
1	月份	2017年氣溫 x	2018年氣溫 y
2	1	1.6	0.5
3	2	0.1	-0.3
4	3	0.3	1.4
5	4	3.4	4.8
6	5	5.4	8
7	6	7.6	8.3
8	7	8.5	7.6
9	8	9.2	7.2
10	9	9.7	7.1
11	10	8.3	4.8
12	11	5	4.4
13	12	2.1	5.1
14	相關係數	0.823583249	

❶選取要製作散布圖的兩行欄位，左邊那一欄為橫軸（x），右邊那一欄為縱軸（y）

❷點擊[插入]標籤。

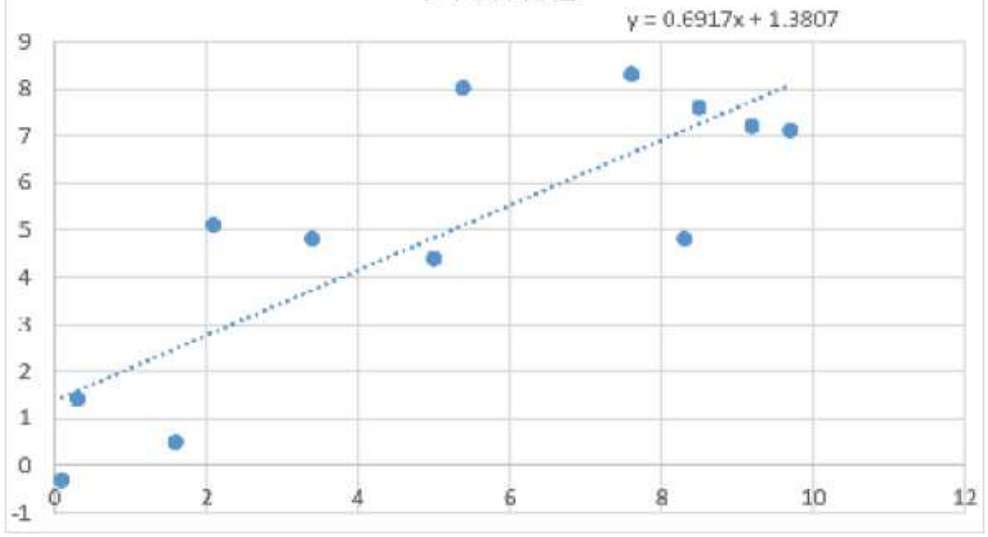
❸從[圖表]群組中，點擊 [散布圖]

❹選擇第一個散布圖。

❺在散布圖的點上，點擊滑鼠右鍵

❻選取[加上趨勢線]。

❼此時螢幕右方會顯示[趨勢線格式]的窗格，分別選取[線性]、[圖表上顯示公式]



故由散布圖中可獲得 y 對 x 的迴歸直線方程式為 $y = 0.6917x + 1.3807$。