

2-2 二維數據分析

重點整理

具有兩個變量的數據，稱為二維數據。

例：小明身高為 176 公分，體重為 72 公斤，用 $(176, 72)$ 表示小明的身高與體重。

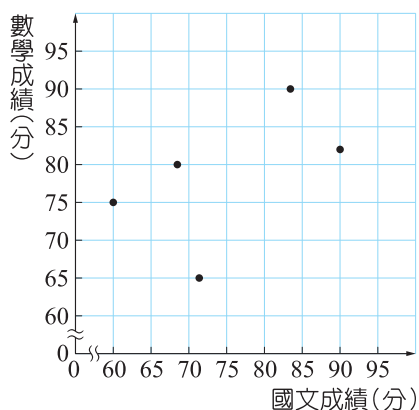
例：這次段考，小華國文 83 分，數學 85 分，用 $(83, 85)$ 表示小華的國文、數學成績。

一、散布圖

1. 以二維數據 (x, y) 當作坐標，將所有二維數據標示在坐標平面上所成的圖形，稱為散布圖。

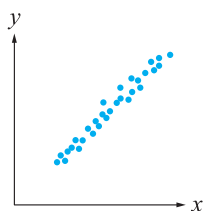
例：這次段考，五位同學的(國文成績，數學成績)為 $(72, 65)$ ， $(83, 90)$ ， $(60, 75)$ ， $(90, 82)$ ， $(68, 80)$ ，試作這五個二維數據的散布圖。

將這五個二維數據當作坐標，標示在坐標平面上，得散布圖如下：

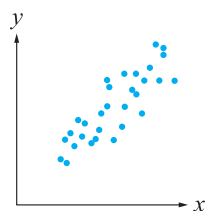


2. 由散布圖觀察兩變量的關係：

- (1) 正相關：兩個變量大約有一致的趨勢(大約同時增加或減少)，如圖(一)、(二)。

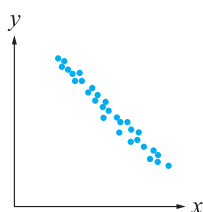


圖(一)

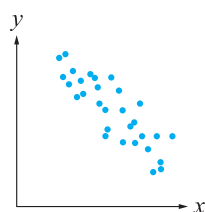


圖(二)

- (2) 負相關：兩個變量趨勢大約相反，一個增加，則另一個大概就會減少；或一個減少，另一個大概就會增加，如圖(三)、(四)。

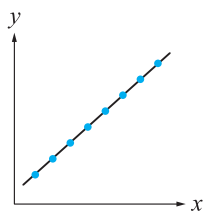


圖(三)

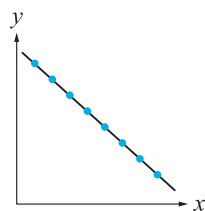


圖(四)

(3) 完全正相關：資料全部在一條斜率為正的直線上，如圖(五)。



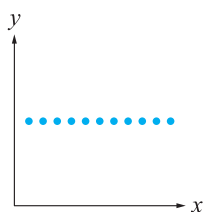
圖(五)



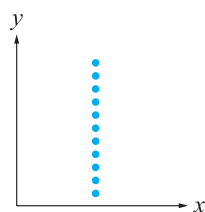
圖(六)

(4) 完全負相關：資料全部在一條斜率為負的直線上，如圖(六)。

(5) 零相關：兩個變量的變化之間無關，如圖(七)、(八)。



圖(七)



圖(八)

二、相關係數

相關係數：用 r 表示，可以判斷兩個變量的相關程度，且 $-1 \leq r \leq 1$ 。

1. 由標準化數據求相關係數：

假設有一群二維數據 (x_i, y_i) ， $i = 1, 2, \dots, n$ ，其中 x_i 的平均數、標準差為 μ_x 、 σ_x ，而 y_i 的平均數、標準差為 μ_y 、 σ_y 。

(1) 先將 x_i, y_i 標準化，得標準化數據 (X_i, Y_i) ，其中 $X_i = \frac{x_i - \mu_x}{\sigma_x}$ ， $Y_i = \frac{y_i - \mu_y}{\sigma_y}$ 。

(2) (x_i, y_i) 的相關係數為 $r = \frac{X_1Y_1 + X_2Y_2 + \dots + X_nY_n}{n}$ 。

2. 由原始數據求相關係數：

原始數據 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ，其中 x_i, y_i 的平均數分別為 μ_x, μ_y ，相關係數為

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \text{ 其中 } S_{xy} = (x_1 - \mu_x)(y_1 - \mu_y) + \dots + (x_n - \mu_x)(y_n - \mu_y),$$

$$S_{xx} = (x_1 - \mu_x)^2 + \dots + (x_n - \mu_x)^2, S_{yy} = (y_1 - \mu_y)^2 + \dots + (y_n - \mu_y)^2。$$

3. 相關係數 r 的性質：

(1) $r > 0$ 表示兩變量為正相關， $r < 0$ 為負相關， $r = 0$ 為零相關。

(2) $-1 \leq r \leq 1$ 。

(3) $r = 1$ 表兩變量為完全正相關， $r = -1$ 為完全負相關。

(4) $|r|$ 愈大表示兩變量的相關程度愈強。

4. 平移伸縮後的相關係數：

已知一群二維數據 (x_i, y_i) 的相關係數為 r ，將 (x_i, y_i) 平移、伸縮為另一群二維數據 (x'_i, y'_i) ，其中 $x'_i = ax_i + b$ ， $y'_i = cy_i + d$ 。

若 a, c 同號 ($ac > 0$)，則 (x'_i, y'_i) 的相關係數 $r' = r$ ；

若 a, c 異號 ($ac < 0$)，則 (x'_i, y'_i) 的相關係數 $r' = -r$ 。

例：已知一群二維數據 (x_i, y_i) ，其中 x_i 與 y_i 的相關係數為0.6，

- (1) 若 $x'_i = 3x_i + 4$ ， $y'_i = 2y_i - 6$ ，因為3與2同號（都是正數），所以相關係數 $r' = 0.6$ 。
- (2) 若 $x'_i = 2x_i + 5$ ， $y'_i = -3y_i + 4$ ，因為2與-3異號（一正一負），所以相關係數 $r' = -0.6$ 。

三、最小平方法與最適直線

用最小平方法找最適直線（又稱迴歸直線）。

高斯等數學家提出的方法，找一條直線 L ，使資料點到 L 的鉛垂距離的平方和最小。

1. 標準化數據的最適直線：

將一群二維數據 (x_i, y_i) 標準化以後，得標準化數據 (X_i, Y_i) 。則最適直線方程式為 $Y = rX$ ，其中 r 為相關係數。

2. 原始二維數據的最適直線：

此線必過 (μ_x, μ_y) ，且斜率為 $r \cdot \frac{\sigma_y}{\sigma_x}$ ，即方程式為

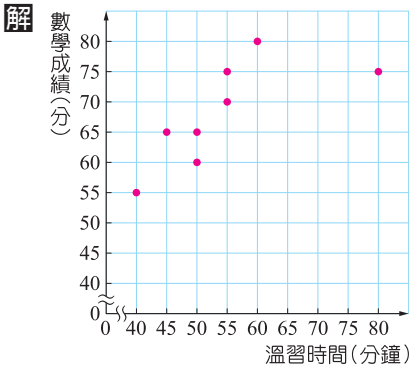
$$y - \mu_y = r \cdot \frac{\sigma_y}{\sigma_x} \cdot (x - \mu_x), \text{ 其中 } r \text{ 為相關係數,}$$

$$\text{或 } y - \mu_y = \frac{S_{xy}}{S_{xx}} \cdot (x - \mu_x)。$$

例題 1 作散布圖(一)(原始數據)

8 位同學每日溫習數學的時間(分鐘)與數學成績(分)如下表，試以時間為 x 坐標，分數為 y 坐標，作散布圖。(10 分)

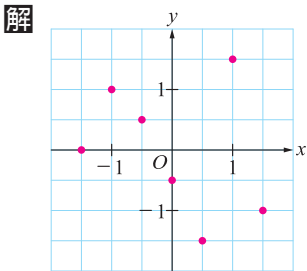
	A	B	C	D	E	F	G	H
溫習時間(分鐘)	50	40	55	55	45	80	50	60
數學成績(分)	65	55	70	75	65	75	60	80



例題 2 作散布圖(二)(標準化數據)

一組二維數據的標準化數據如下表，試作散布圖。(10 分)

x	-0.5	0.5	1.5	1	0	-1.5	-1
y	0.5	-1.5	-1	1.5	-0.5	0	1



例題 3 計算相關係數(一)(標準化數據)

一組標準化的二維數據如下表，試求相關係數。(10分)

X	1.5	0	0.5	-1.5	-0.5
Y	-0.5	-1.5	1.5	0	0.5

解 將同一筆資料的 X 與 Y 相乘，相加以後，除以 5 即可

X	1.5	0	0.5	-1.5	-0.5
Y	-0.5	-1.5	1.5	0	0.5
XY	-0.75	0	0.75	0	-0.25

$$\text{故相關係數 } r = \frac{(-0.75) + 0 + 0.75 + 0 + (-0.25)}{5} = \frac{-0.25}{5} = -0.05$$

例題 4 計算相關係數(二)(原始數據)

一組二維數據如下表，試求相關係數。(10分)

x	32	23	26	14	20
y	15	13	19	7	11

解 計算 x 與 y 的平均數，得 $\mu_x = 23$ ， $\mu_y = 13$

x	y	$x - \mu_x$	$y - \mu_y$	$(x - \mu_x)(y - \mu_y)$	$(x - \mu_x)^2$	$(y - \mu_y)^2$
32	15	9	2	18	81	4
23	13	0	0	0	0	0
26	19	3	6	18	9	36
14	7	-9	-6	54	81	36
20	11	-3	-2	6	9	4
總和				96	180	80

$$\text{故得相關係數 } r = \frac{96}{\sqrt{180} \times \sqrt{80}} = \frac{96}{6\sqrt{5} \times 4\sqrt{5}} = 0.8$$

例題 5 計算相關係數(≡)(平移伸縮)

已知一群二維數據 (x_i, y_i) 的相關係數為 0.45，將 (x_i, y_i) 平移伸縮為新的數據 (x'_i, y'_i) ，

(1) 若 $(x'_i, y'_i) = (5x_i + 1, -6y_i - 7)$ ，試求相關係數 r' 。(5 分)

(2) 若 $(x''_i, y''_i) = (-5x_i + 30, -2y_i + 25)$ ，試求相關係數 r'' 。(5 分)

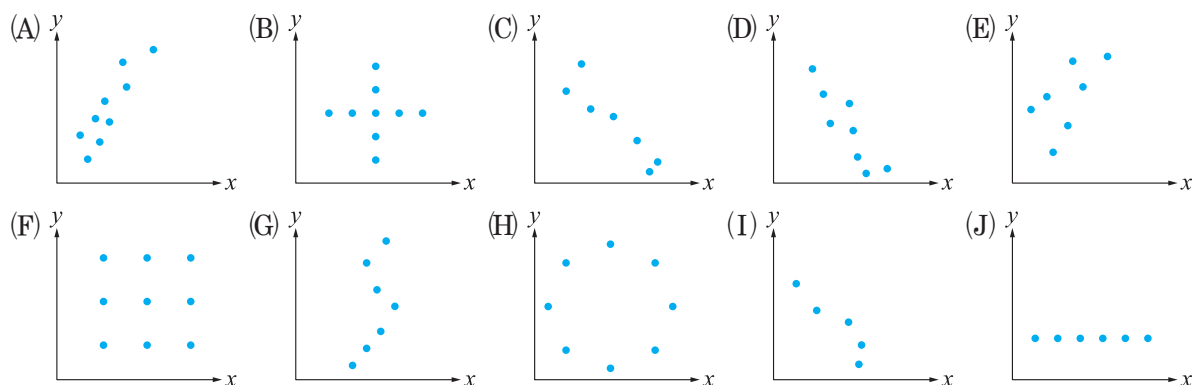
解 二維數據平移伸縮以後的相關係數，只需觀察伸縮倍數的正負號即可

(1) $\because 5$ 與 -6 異號 $\therefore r' = -0.45$

(2) $\because -5$ 與 -2 同號 $\therefore r'' = 0.45$

例題 6 判斷相關係數(散布圖與相關係數)

觀察下列 10 個散布圖，判斷其相關係數的正負，請選出符合下列條件的散布圖：



(1) 相關係數大於 0。(5 分)

(2) 相關係數小於 0。(5 分)

解 (1) 圖(A)(E)(G)資料的分布靠近一條斜率大於 0 的直線

表示兩個變量的相關係數大於 0

(2) 圖(C)(D)(I)資料的分布靠近一條斜率小於 0 的直線

表示兩個變量的相關係數小於 0

其餘(B)(F)(H)(J)資料的分布呈對稱

表示兩個變量的相關係數等於 0

例題 7 求最適直線(一)(用標準化數據計算)

一組二維數據的標準化數據如下表，試求最適直線的方程式。(10分)

X	1.5	0.5	0	-1.5	-0.5
Y	0.5	0	1.5	-0.5	-1.5

解 標準化數據的最適直線方程式為 $Y = rX$ ，故須求出此兩變量的相關係數
將同一筆資料的 X 與 Y 相乘，相加以後，除以 5 即可

X	1.5	0.5	0	-1.5	-0.5
Y	0.5	0	1.5	-0.5	-1.5
XY	0.75	0	0	0.75	0.75

$$\text{相關係數 } r = \frac{0.75 + 0 + 0 + 0.75 + 0.75}{5} = 0.45$$

故得最適直線為 $Y = 0.45X$

例題 8 求最適直線(二)(用原始數據計算)

一組二維數據如下表，試求最適直線的方程式。(10分)

x	46	43	52	55	49
y	24	32	48	44	32

解 計算 x 與 y 的平均數，得 $\mu_x = 49$ ， $\mu_y = 36$

x	y	$x - \mu_x$	$y - \mu_y$	$(x - \mu_x)(y - \mu_y)$	$(x - \mu_x)^2$
46	24	-3	-12	36	9
43	32	-6	-4	24	36
52	48	3	12	36	9
55	44	6	8	48	36
49	32	0	-4	0	0
總和				144	90

得最適直線斜率為 $\frac{144}{90} = \frac{8}{5}$ ，故最適直線方程式為 $y - 36 = \frac{8}{5}(x - 49)$

例題 9 求最適直線(已知平均數、標準差與相關係數)

假設高一某班第一次段考的國文、數學成績以 (x_i, y_i) 表示，平均數分別為 $\mu_x = 65$ 、 $\mu_y = 70$ ，標準差分別為 $\sigma_x = 8$ 、 $\sigma_y = 12$ ，兩科成績的相關係數為 0.6。試求：

- (1) 數學成績對國文成績的最適直線方程式。(5 分)
- (2) 已知小明的國文成績 85 分，試推估小明的數學成績。(5 分)

解 (1) 最適直線必通過 (μ_x, μ_y) ，故此線通過點 $(65, 70)$

$$\text{斜率 } m = 0.6 \times \frac{12}{8} = 0.9$$

故最適直線方程式為 $y - 70 = 0.9(x - 65)$ ，即 $y = 70 + 0.9(x - 65)$

- (2) 令 $x = 85$ 代入最適直線方程式，得 $y = 70 + 0.9(85 - 65) = 88$ (分)

例題 10 求最適直線並作預估

根據某商店的銷售紀錄，7 月 1 日開始，連續 5 日每日最高溫與當日冰棒銷售量如下表：

日期	7/1	7/2	7/3	7/4	7/5
最高溫($^{\circ}\text{C}$)	30	34	36	33	32
銷售量(枝)	430	500	600	540	560

- (1) 試以每日最高溫為 x 坐標，每日銷售量為 y 坐標，求最適直線方程式。(5 分)
- (2) 若氣象局預測第 6 日最高溫為 35°C ，試以(1)的結果，預估第 6 日的冰棒銷售量。(5 分)

解 (1) 計算 x 與 y 的平均數，得 $\mu_x = 33$ ， $\mu_y = 526$

x	y	$x - \mu_x$	$y - \mu_y$	$(x - \mu_x)(y - \mu_y)$	$(x - \mu_x)^2$
30	430	-3	-96	288	9
34	500	1	-26	-26	1
36	600	3	74	222	9
33	540	0	14	0	0
32	560	-1	34	-34	1
總和				450	20

$$\text{得最適直線斜率為 } \frac{450}{20} = \frac{45}{2}$$

$$\text{故最適直線方程式為 } y - 526 = \frac{45}{2}(x - 33), \text{ 即 } y = 526 + \frac{45}{2}(x - 33)$$

- (2) 令 $x = 35$ 代入(1)的最適直線方程式，得 $y = 571$ ，故預估第 6 日銷售量為 571 枝