素養導向高級中學數學教材

相關係數與最佳直線

國家教育研究院

十二年國民基本教育數學領域教材與教學模式研發編輯小組

最佳直線與相關係數

壹 歷史與生活

1 歷史

法蘭西斯・高爾頓爵士 *Sir Francis Galton*
（西元1822～1911，英國人）

1884 年，英國人類學家高爾頓（*Sir Francis Galton*）在倫敦成立人體測量實驗室，收集了許多關於親子間的資料，包括身高、體重、特定骨頭的長度等。他發現「非常高的父母所生的孩子，往往會比父母矮些，而非常矮的父母所生的孩子，則往往比父母高」，他把這個現象稱作「迴歸至平均值（*regression to the mean*）」，這就是現在的統計上「迴歸（*regression*）」一詞的起源。

事實上，高爾頓是演化論之父達爾文（*Charles Darwin*）的表弟，高爾頓原本在劍橋讀醫學。1860 年時，高爾頓轉向氣象學的研究，在這段研究過程中，他對於統計方面的興趣與能力漸漸的浮現。1865 年起，高爾頓由於自己家族的經驗以及達爾文的影響，興趣轉向於人類種族的進化與遺傳學，並開始了研究統計上的問題。而他最為大家熟知的事蹟，便是首先發現了不同人、不同種族具有不同指紋。這讓人們知道世界上每個人的指紋都是獨一無二的，甚至有特定的方法可用來區分並辨識一個人的身分。

起初，為了瞭解遺傳的特性，高爾頓試圖從智力演化的方向去探討，卻礙於當時沒有一套完善測量智力的方法而遭遇到瓶頸。於是他想到一個能容易測量且公正的人類特徵「身高」，這才有了人體測量實驗室的成立。

高爾頓在達爾文的《物種原始》一書中提到他對於遺傳的看法：「這些新的觀念，激勵我去研究遺傳學和人類種族的進化。」此時，他需要一個好方法來描述這個世代的智力，與前一個世代的智力是「相關」的。高爾頓再嘗試尋找可供測量如此關係的數學方法後，他開始使用了相關係數（*correlation coefficient*）的概念。他使用字母「*r*」來表示相關係數，而這個傳統一直延續至今。現今的相關係數的公式是由高爾頓的學生皮爾森（*Karl Pearson*）所發展出來的。

（資料來源：國立臺灣大學「生物統計學程」*http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm#gal*）

2 生活

在網路新聞上搜尋「相關係數」一詞，可發現它在經濟、科學、政治等生活應用的各種新聞不少，例如：

『生活幸福感是一個非常主觀的概念，在這一次的調查中，我們針對「生活幸福感」作出調查，同時透過和生活幸福感可能有關的 11 個面向分別進行電話訪問。調查結束之後，統計分析顯示，按照相關程度的高低，和生活幸福感最相關的面向分別為：未來發展樂觀度（相關係數為 0.545）、經濟收入（0.457）、工作情況（0.450）、家庭關係（0.362）、人際關係（0.319）、地方政府施政（0.291）、環境品質（0.276）、健康狀況（0.270）、政治權利（0.265）。至於治安狀況則不具有統計解釋力、宗教信仰相關係數偏低，這兩個面向因此只有表面上的參考價值，我們不再作深入的探討。』（2012/05/17　幸福指數的重要性 臺灣競爭力論壇彭錦鵬，臺灣競爭力論壇理事長）

甚至，我們會在財經新聞上聽到這樣的報導：「歷史經驗顯示，美國聯準會升息前，美元會有一波明顯上漲的走勢，而美國十年期公債與基準利率相關係數高達 0.92（呈高度正相關），且殖利率曲線走勢明顯快於聯邦基準利率，因此，可視這兩指標為美國何時升息的領先指標……。」（2015/07/31　從 7 月的利率會議聲明，來看 9 月美聯儲升息的機率！）

現在生活周遭中許多變數間關聯性的探討，與種種分析數據的方法，其實是源自於數百年來科學家們努力的成果。

貳 直線相關

1 散布圖與相關

前言：

在日常生活中，我們也常常將兩個數據資料相提並論，例如：吸菸與肺癌、咖啡因與骨質疏鬆症、睡眠時數與肥胖程度、國民所得與壽命、產品的售價與需求量等等。

針對兩個數據資料之間可以討論以下三個問題：

⑴ 兩個數據資料間的關聯性為何？

⑵ 如何衡量兩數據資料直線相關的程度？

⑶ 如何找出最佳的直線來描述兩數據資料的關係並作預測？

活動 1

數學成績高的學生，物理成績通常也不會很低嗎？

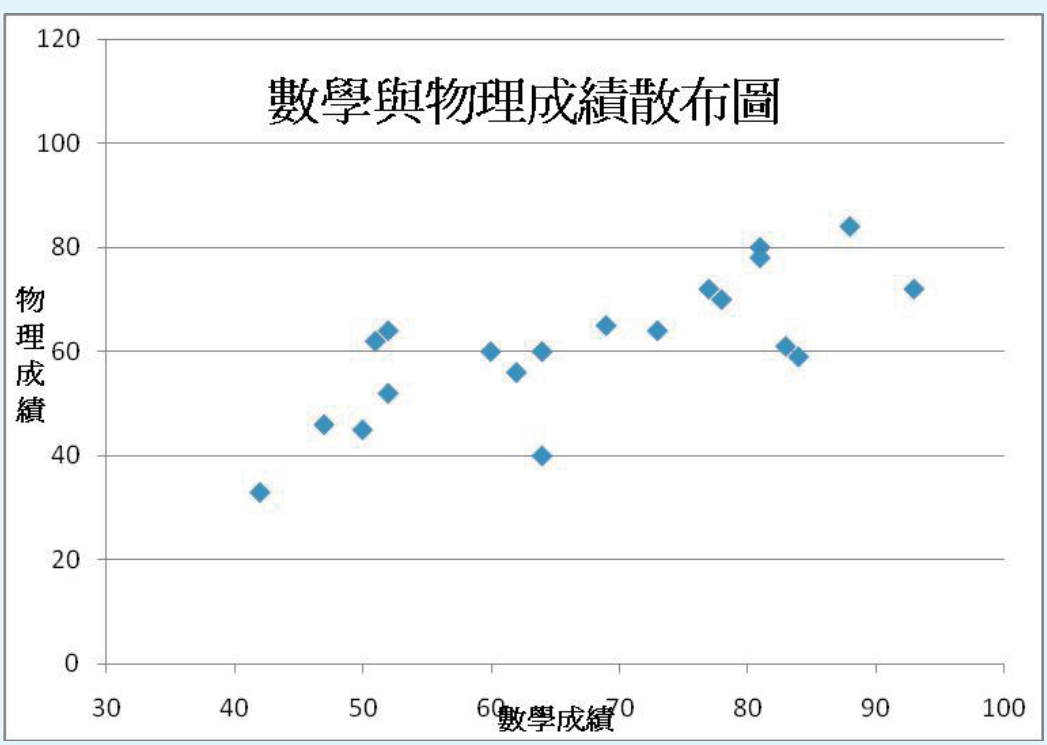
考慮某個社團中成員數學與物理的成績：

編號	1	2	3	4	5	6	7	8	9	10
數學	93	88	83	42	50	81	52	69	73	78
物理	72	84	61	33	45	80	64	65	64	70

編號	11	12	13	14	15	16	17	18	19	20
數學	84	52	77	64	62	64	60	81	51	47
物理	59	52	72	40	56	60	60	78	62	46

將兩個數據資料，以數對方式畫在坐標平面上，以表明它們的分布情形的圖形，如圖所示，稱為**散布圖**，散布圖上的點稱為**樣本點**。

觀察數學與物理的散布圖，經由計算數學與物理成績的平均數分別為67.55 與61.15 分，是否有數學成績超過（低於）平均數，而物理成績超過（低於）平均數的趨勢？



活動 2

適量的飲用葡萄酒可以預防心臟病？

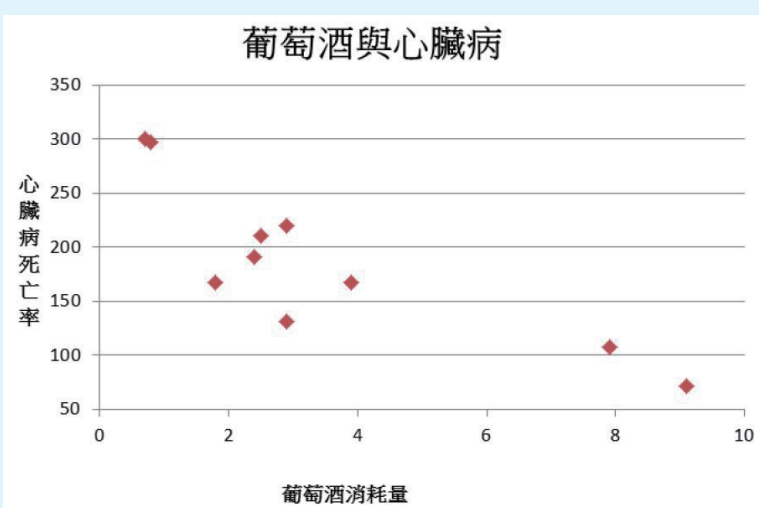
下表是 10 個已開發國家一年葡萄酒消耗量（平均每人從喝葡萄酒所攝取的酒精量）以及一年中因心臟病死亡率（每十萬人死亡人數）。

國家	澳洲	奧地利	比利時／盧森堡	加拿大	丹麥
葡萄酒消耗量（公升）	2.5	3.9	2.9	2.4	2.9
心臟病死亡率（每十萬人死亡人數）	211	167	131	191	220

國家	芬蘭	法國	荷蘭	愛爾蘭	義大利
葡萄酒消耗量（公升）	0.8	9.1	1.8	0.7	7.9
心臟病死亡率（每十萬人死亡人數）	297	71	167	300	107

※ 資料來源出自《統計學的世界》*P*400（*David S.Moore*著，鄭惟厚譯，天下文化）

觀察上述資料的散布圖，經由計算葡萄酒消耗量、心臟病死亡率的平均數分別為 3.49 公升、186.2 人，是否有葡萄酒消耗量超過（低於）平均數的國家，他們人民心臟病死亡率低於（高於）平均數的趨勢？



根據前面兩個問題，可以得到以下結論：

1. 散布圖（*scatter plot*）的意義：

蒐集了兩數據資料 X 與 Y：(x_1, y_1)、(x_2, y_2)、……、(x_n, y_n)，將每一個數對 (x_i, y_i) 標示在坐標平面上，所得的圖形稱為**散布圖**，散布圖上的點稱為**樣本點**。

從散布圖中，我們可以觀察資料分布的整體型態與相關情形。

2. 正相關與負相關

從散布圖中觀察兩個數據資料 X 與 Y 之間的相關情形，當其中一個數據的值高於平均時，另一數據的值也傾向高於平均；而其中一個數據的值低於平均時，另一數據的值也傾向低於平均，則稱數據資料 X 與 Y 是**正相關（positively associated），此時樣本點大致上會從左下往右上傾斜**。如果其中一個數據的值高於平均時，另一數據的值傾向低於平均；而若其中一個數據的值低於平均時，另一數據的值傾向高於平均，則稱數據資料 X 與 Y 是**負相關（negatively associated），此時樣本點大致上會從左上往右下傾斜**。

任務 1

從散布圖判別正相關與負相關：

⑴活動 1 中，請問兩筆資料是正相關或是負相關？

⑵活動 2 中，請問兩筆資料是正相關或是負相關？

任務 2

某公司在過去數年擁有穩定的月銷售金額。今年該公司決定調整廣告費用以觀察是否對銷售金額有明顯影響，為了利於評估廣告對銷售金額的影響，該公司蒐集的資料數據如下表。其中每月廣告費用 x_i（單位：千元）與銷售金額 y_i（單位：千元）。

月分	1	2	3	4	5	6	7	8	9	10
廣告費用 x_i（千元）	2	4	6	5	3	5	4	3	5	7
銷售金額 y_i（千元）	35	50	60	60	45	55	40	40	50	65

⑴請根據上述資料畫出散布圖，並將「銷售金額」置於垂直坐標軸上。

⑵ 請描述資料分布的整體型態及「廣告費用」、「銷售金額」二者的關聯性。

任務 3

某肥皂廠商欲推出一種新產品，在上市前以不同的單價 x （單位：十元）調查市場的需求 y （單位：萬盒），調查結果如下表：

單價 x	8	9	10	11	12	13
需求 y	13	12.4	11	9.6	9.2	8

⑴ 請根據上述資料畫出散布圖，並將「需求」置於垂直坐標軸上。

⑵ 請描述資料分布的整體型態及「單價」、「需求」二者的關聯性。

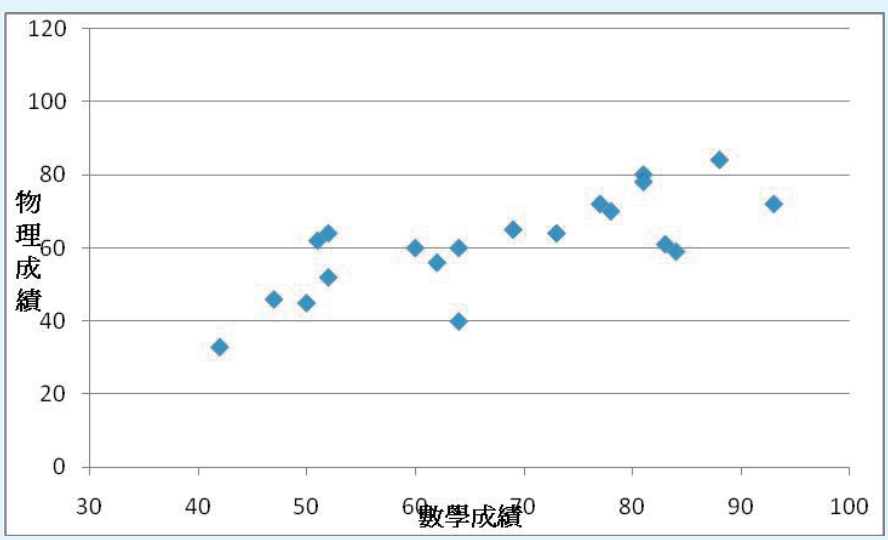
叁 最佳直線與相關係數

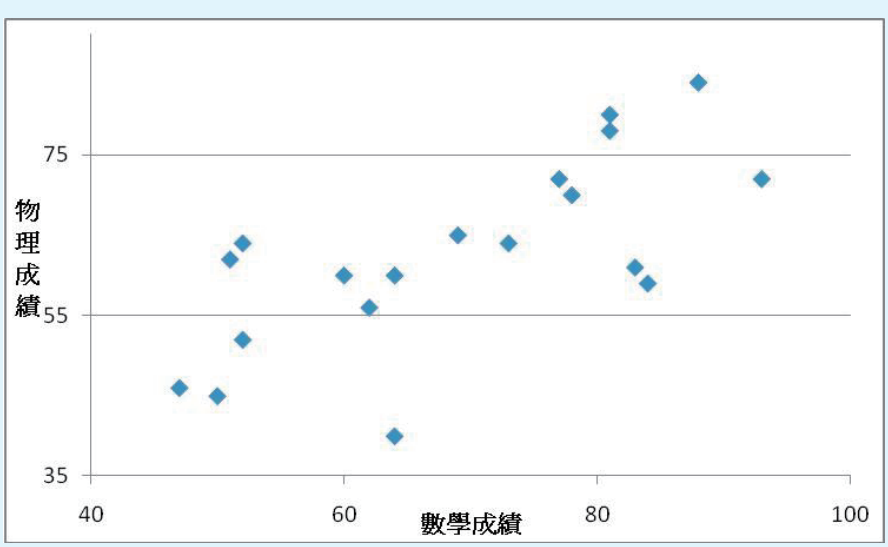
1 最小平方法

散布圖呈現兩個數據資料間相關的方向、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。

活動 3

只靠散布圖判別兩變量的相關足夠嗎？





上面兩個散布圖，哪個圖的數學、物理成績直線相關比較強？

活動三中，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，所以只要修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。所以得定義一個統計量（相關係數）來衡量兩個變數的直線相關強度，我們先從代表兩筆數據的直線開始，探討如何找出最佳（最適合）直線並定義相關係數。

活動 4 最小平方法的引進

右表中有 4 筆資料：

x	1	2	3	4
y	3	1	2	7

若想用直線 $y=a+bx$ 來表示 x、y 的關係，那麼 a、b 要如何取，才會使直線 $y=a+bx$ 與散布圖中的點愈靠「近」呢？

⑴令樣本點 $(x_1, y_1)=(1, 3)$、$(x_2, y_2)=(2, 1)$、$(x_3, y_3)=(3, 2)$、$(x_4, y_4)=(4, 7)$，希望能夠選取 a、b 的值，使得資料點 x_i 的 y 坐標 y_i（實際值）與 $a+bx_i$（預測值）的誤差要最小。

請就下面幾種誤差的形式加以討論它們有甚麼優缺點。

$$E_1=|(y_1-(a+bx_1))+(y_2-(a+bx_2))+(y_3-(a+bx_3))+(y_4-(a+bx_4))|$$

$$E_2=|y_1-(a+bx_1)|+|y_2-(a+bx_2)|+|y_3-(a+bx_3)|+|y_4-(a+bx_4)|$$

$$E_3=(y_1-(a+bx_1))^2+(y_2-(a+bx_2))^2+(y_3-(a+bx_3))^2+(y_4-(a+bx_4))^2$$

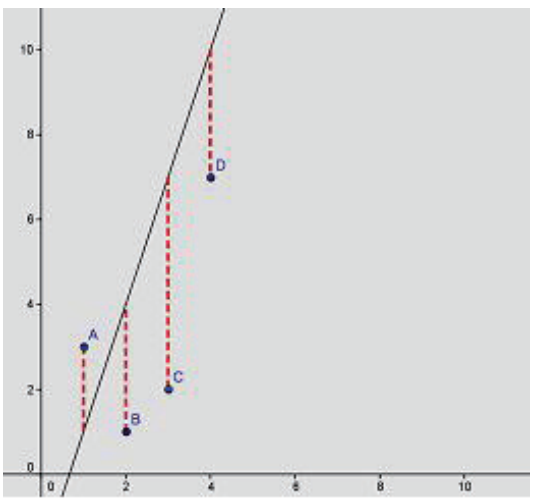
$$E_4=\frac{1}{4}[(y_1-(a+bx_1))^2+(y_2-(a+bx_2))^2+(y_3-(a+bx_3))^2+(y_4-(a+bx_4))^2]$$

⑵經計算 x、y 兩筆數據資料的算術平均數分別為 $\mu_x=\frac{10}{4}$，$\mu_y=\frac{13}{4}$。考慮通過 (μ_x, μ_y) 斜率為 m 的直線 $y=m(x-\mu_x)+\mu_y$，利用 GeoGebra 軟體找出誤差最小時斜率為 m 的值。

如果散布圖顯示出兩個數值資料之間有很強的直線相關，可以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，找出來的直線稱為**最佳直線**或**迴歸直線**。

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)、求出一條直線$y=a+bx$使得誤差平方的平均$E=\frac{1}{n}(\sum_{i=1}^{n}〔y_i-(a+bx_i)〕^2)$最小。
求得的直線$y=a+bx$稱為**y對x的最佳直線或迴歸直線**。



活動 5　用最小平方法找最佳直線

考慮活動 4 中的 4 個樣本點：
$(x_1, y_1)=(1, 3)$、$(x_2, y_2)=(2, 1)$、$(x_3, y_3)=(3, 2)$、$(x_4, y_4)=(4, 7)$，
根據配方法，找出a、b使得誤差

$$E=\frac{1}{4}〔(3-(a+b))^2+(1-(a+2b))^2+(2-(a+3b)^2+(7-(a+4b))^2〕$$

$$=\frac{1}{4}〔(2a+5b-\frac{13}{2})^2+5(b-\frac{13}{10})^2+\frac{123}{10}〕\text{最小。}$$

任務 4　考慮 3 個樣本點

$(x_1, y_1)=(1, 2)$、$(x_2, y_2)=(2, 1)$、$(x_3, y_3)=(3, 3)$，求兩實數a、b使得下列E值最小：

$$E=\frac{1}{3}〔(y_1-a-bx_1)^2+(y_2-a-bx_2)^2+(y_3-a-bx_3)^2〕$$

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

$$=\frac{1}{3}〔3(a+2b-2)^2+2(b-\frac{1}{2})^2+\frac{3}{2}〕$$，試求此兩筆數據資料的最佳直線。

2 定義相關係數

活動 6

如下表所示，給定 X、Y 兩個數據資料，

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

若 X 與 Y 的關係可以用直線來描述，利用最小平方法可以找到 Y 對 X 的最佳直線 L：$y=a+bx$ 使得誤差 $E=\frac{1}{n}\left(\sum_{i=1}^{n}\left[y_i-(a+bx_i)\right]^2\right)$ 最小。

為了配方方便起見，將 X、Y 兩個數據資料標準化成 X'、Y'

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

其中 $x_i'=\frac{x_i-\mu_x}{\sigma_x}$，$y_i'=\frac{y_i-\mu_y}{\sigma_y}$。

設標準化後，由最小平方法得到 Y' 對 X' 的最佳直線 L'：$y'=a+bx'$

考慮誤差 $E'=\frac{1}{n}\sum_{i=1}^{n}\left[y_i'-(a+bx_i')\right]^2$

⑴數據資料 X'、Y' 的平均數與標準差分別為 0 與 1，

試求下列各項之值 $\sum_{i=1}^{n}x_i'$，$\sum_{i=1}^{n}y_i'$，$\sum_{i=1}^{n}(x_i')^2$，$\sum_{i=1}^{n}(y_i')^2$。

⑵誤差 $E'=\frac{1}{n}\sum_{i=1}^{n}\left[y_i'-(a+bx_i')\right]^2$ 可以配方化成

$a^2+\left[b-\frac{1}{n}\sum_{i=1}^{n}(x_i'y_i')\right]^2+1-\left[\frac{1}{n}\sum_{i=1}^{n}(x_i'y_i')\right]^2$

根據上式可以得知當 a、b 之值為何時，E' 有最小值？

⑶試求數據資料標準化之後的最佳直線。

活動六中，令 $r=\frac{1}{n}\sum_{i=1}^{n}(x_i' y_i')$，接下來，我們來討論 r 的範圍，以及 r 與最佳直線的關係。

活動 7

活動六中，令 $r=\frac{1}{n}\sum_{i=1}^{n}(x_i' y_i')$

⑴請問 Y' 對 X' 的最佳直線如何表示？（以 r 表示）

⑵請問 r 的範圍為何？

⑶請討論當 r 改變時，選用最佳直線代表數據資料的關係是否合適？

根據活動六、七的討論，$r=\frac{1}{n}\sum_{i=1}^{n}(x_i' y_i')$ 可以作為衡量兩個變數 X、Y 直線相關的強弱程度的統計量，我們稱為**相關係數**。

※ 相關係數（correlation coefficient）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：

對於兩組數據資料 X、Y

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

X 與 Y 的相關係數 r 定義為 $\frac{1}{n}\sum_{i=1}^{n}(x_i' y_i')$，

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

相關係數亦可以寫成

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

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。

任務 5

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

根據前面的討論，相關係數 r 可以量測兩個數據資料的直線相關程度，並且具有以下特性：

⑴它的範圍有界，強度大小的絕對值不大於 1，即 $-1\le r\le 1$。

⑵它能表達出相關性的正負方向。

⑶它與變數所使用的量測單位無關。

⑷它能表達出兩變數間直線相關性的強度大小。

說明如下：

⑴相關係數的範圍有界，強度大小的絕對值不大於 1，

即 $-1\le r\le 1$。根據活動六的結果即可得知。

⑵相關係數能表達出直線相關的方向。

設 (X, Y) 的數據資料為 (x_1, y_1)、(x_2, y_2)、…、(x_n, y_n)，在散布圖中以 $y=\mu_y$ 為新的橫軸，$x=\mu_x$ 為新的縱軸，則可將散布圖分成四個象限，

如果點 (x_i, y_i) 在第一、三象限內，則 $(x_i-\mu_x)(y_i-\mu_y)$ 的值為正；

如果點 (x_i, y_i) 在第二、四象限內，則 $(x_i-\mu_x)(y_i-\mu_y)$ 的值為負，因此

(*a*) 若 $r>0$ 時，即 $\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)>0$，則 X、Y 為正相關。

即表示 X 與 Y 的變動趨勢大致相同。

(*b*) 若 $r<0$ 時，即 $\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)<0$，則 X、Y 為負相關。

即表示 X 與 Y 的變動趨勢大致相反，即此增彼減或此減彼增。

⑶相關係數與數據所使用的量測單位無關。

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

分子、分母單位相消，所以相關係數 r 與使用的單位無關。

⑷相關係數能表達出直線相關的強度。

(*a*) 若 r 值很接近 0，表示兩變數之間有很弱的直線相關；

(*b*) 若 r 的絕對值越接近 1，表示兩變數之間的直線相關程度越大。

(*c*) $r=1$ 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。

3 找最佳直線

根據前面的討論，對於標準化的資料 X'、Y' 而言，Y' 對於 X' 的最佳直線為 $y'=rx'$。因為 $x_i'=\frac{x_i-\mu_x}{\sigma_x}$，$y_i'=\frac{y_i-\mu_y}{\sigma_y}$，因此可以令 $x'=\frac{x-\mu_x}{\sigma_x}$，$y'=\frac{y-\mu_y}{\sigma_y}$ 代入 $y'=rx'$ 得到 $\frac{y-\mu_y}{\sigma_y}=r\left(\frac{x-\mu_x}{\sigma_x}\right)$，化簡為 $y=\left(r\frac{\sigma_y}{\sigma_x}\right)(x-\mu_x)+\mu_y$。

我們稱 $y=\left(r\frac{\sigma_y}{\sigma_x}\right)(x-\mu_x)+\mu_y$ 為數據資料 Y 對 X 的最佳直線。

任務 6

請問 Y 對 X 的最佳直線的斜率等於多少？一定會通過哪一點？

任務 7

利用 σ_x、σ_y、r 的定義，試推導 Y 對 X 的最佳直線的斜率

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

根據前面的討論，可以整理成以下的結論：

⑴若給定 X、Y 兩筆數據資料，將 X、Y 標準化成數據 X'、Y'，

則 Y' 對 X' 的最佳直線 L' 為 $y'=rx'$，其中 r 為數據 X、Y 的相關係數。

⑵若給定 X、Y 兩筆數據資料，$\begin{array}{c|c|c|c|c} X & x_1 & x_2 & \cdots & x_n \\ \hline Y & y_1 & y_2 & \cdots & y_n \end{array}$，則 Y 對 X 的最佳直線 L：$y=a+bx$ 必通過點（μ_x，μ_y），斜率 $b=\dfrac{r\sigma_y}{\sigma_x}=\dfrac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sum_{i=1}^{n}(x_i-\mu_x)^2}$。

※ 線性模型

要預測必須先有模型，若我們決定模型為線性模型，然後根據蒐集到的數據，利用最小平方法決定直線的斜率和截距，找出最佳直線或迴歸直線。若是兩個數據之間的關係是可以解釋或預測的話，我們就可以透過最佳直線用一個變數來解釋或預測另一個變數。

活動 8

假設臺灣某珍珠奶茶店的店長注意到每天珍珠奶茶的銷售金額似乎與當天的最高氣溫有關，於是隨機選了 6 天並記錄了該日最高氣溫（攝氏）和珍珠奶茶的銷售金額（千元）如下表：

編號	1	2	3	4	5	6	平均
最高氣溫（攝氏 x 度）	31	35	33	37	34	34	34
銷售金額（y 千元）	60	78	81	102	90	75	81

店長觀察數據之後認為「銷售金額」與「最高氣溫」二者之間似乎有某種關聯性，他希望能找到這項關聯，並加以利用，但是他沒有學過數據分析，我們來幫他做這件事吧！

⑴請根據上述資料畫出散布圖。

⑵請描述資料分布的整體型態 及「最高氣溫」、「銷售金額」二者的關聯性。

⑶計算「最高氣溫」、「銷售金額」二者的相關係數。

⑷求這 6 筆資料「最高氣溫」對「銷售金額」的最佳直線方程式。

⑸用最佳直線方程式估計最高氣溫 36 度時珍珠奶茶的銷售金額。

⑹若最高氣溫為攝氏 100 度，則銷售金額為多少元？這樣的預測合理嗎？

任務 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：

$\mu_x=65$，$\mu_y=70$，$\sigma_x=10$，$\sigma_y=5$，$r=0.8$

⑴請寫出英文成績（y）對數學成績（x）的最佳直線方程式。

⑵若此班某位同學數學成績 65 分，請預測此生的英文成績。

任務 9

在活動八中，若店長為提供想加盟開店的美國友人資料，將攝氏溫度（x 度）及臺幣（y 千元）單位分別轉換成華氏溫度（u 度）及美元（v 千美元）。那麼

⑴相關係數會怎麼改變？

⑵以最小平方法決定的最佳直線斜率會怎麼改變？

⑶最佳直線方程式為何？

（已知當攝氏溫度為 x 時，華氏溫度為 $u=\frac{9}{5}x+32$；1 美元以 30 元臺幣計算）

你可以利用 *Excel* 來計算以上各問題。

任務 10

同學們可將任務九的問題一般化：

設有兩個變數 X、Y 的 n 筆數據資料（x_i, y_i），$i=1,2,\cdots,n$。已知 X、Y 的算術平均數分別為 μ_x、μ_y，標準差分別為 σ_x、σ_y，相關係數為 r_{XY}，Y 對 X 的最佳直線斜率為 m。若將變數 X 與 Y 經由伸縮、平移分別得到變數 U、V，其中 $U=aX+b$，$V=cY+d$，其中 a、b、c、d 均為常數，即 $\mu_i=ax_i+b$，$i=1,2,\cdots,n$；$v_i=cy_i+d$，$i=1,2,\cdots,n$。那麼

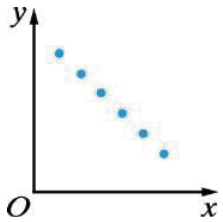
⑴ U、V 的相關係數會怎麼改變？

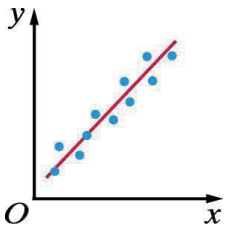
⑵ V 對 U 的最佳直線斜率會怎麼改變？

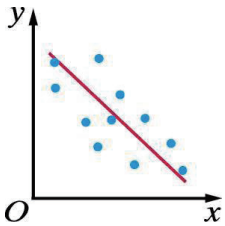
⑶ V 對 U 最佳直線方程式為何？

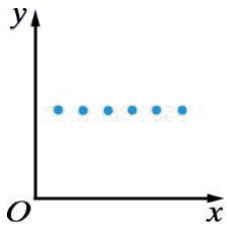
評　量

1.

(A) 

(B) 

(C) 

(D) 

⑴請排出上面 4 個散布圖中 x，y 的相關強度的大小次序（由強到弱）。

⑵請排出上面 4 個散布圖中 x，y 的相關係數的大小次序（由大到小）。

2. 汽車每公升汽油跑的公里數在速度增加時會先上升再下降，假設這種關聯相當規則，汽車行駛的速度（每小時公里數）和汽油里程（每公升公里數）資料所示：

速度	32	48	64	80	96
汽油里程	10.14	11.83	12.68	11.83	10.14

⑴請計算「速度」、「汽油里程」的相關係數。

⑵請解釋為何「速度」、「汽油里程」二者的關聯性很強，但相關係數卻是 0。

3. 請利用下面的數據畫一個散布圖。

x	1	2	3	4	10	10
y	2	4	4	6	2	12

計算相關係數的結果大約是 0.5。對這組數據中的大部分的點來說，x 和 y 之間有很強的直線關聯，是什麼因素導致相關係數只有 0.5 左右？

4. 蒐集學生十人（甲、乙、…、癸），記錄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	4	3	3	4	3	5	4	3	0
成績	100	90	90	80	70	70	60	60	80	100

⑴試求出缺課數與數學成績的相關係數。

⑵設缺課數為 x，數學成績為 y，試求數學成績對缺課數的最佳直線。

⑶若阿杰缺了 10 堂課，根據最佳直線的預測，他的數學成績為多少分？

⑷當缺課數 42 節時，是否仍可以此直線來預測學生的成績？

5. 調查某國家某一年 5 個地區的香煙與肺癌之相關性，所得到的數據為（x_i, y_i），$i=1$、2、3、4、5，其中變數 X 表示每人每年香煙消費量（單位：十包），Y 表示每十萬人死於肺癌的人數。

若已計算出下列數值

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

求 X 與 Y 的相關係數。

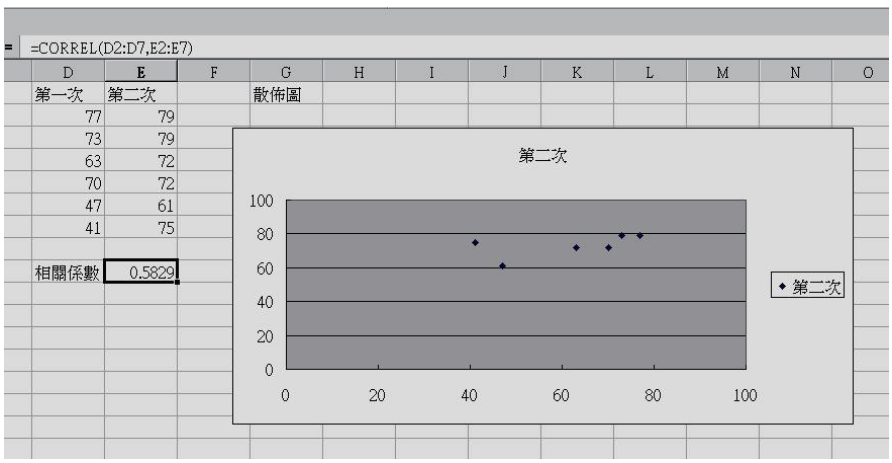
6. 高爾頓（*Galton*）當時曾研究過肘長與身高的相關性，我們可以找幾位同學，測量其肘長與身高，畫出散布圖。

並判定肘長與身高兩數據資料是正相關或是負相關？

計算相關係數與求身高（y）對肘長（x）的最佳直線，並利用預測同學身高看看準不準？

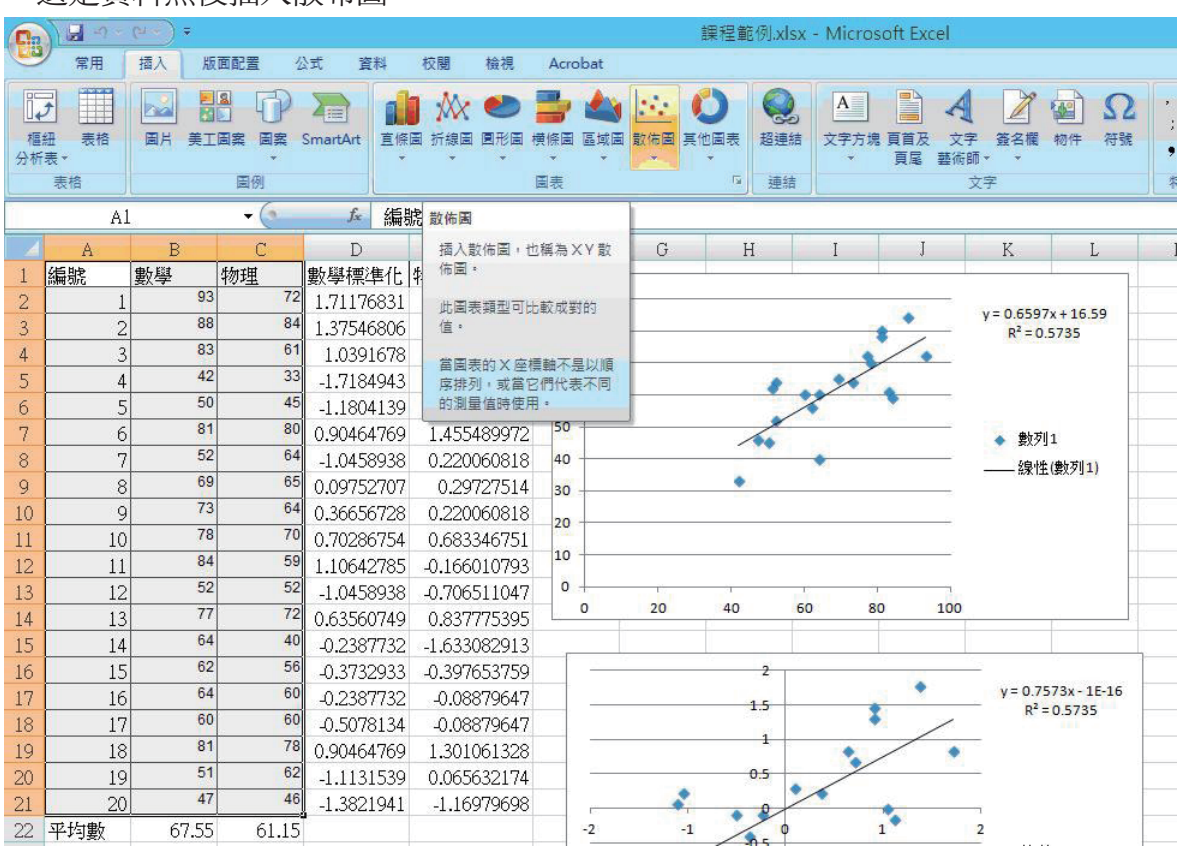
附錄一

1. 用 *Excel* 計算相關係數：



2. 利用 *Excel* 畫散布圖

選定資料然後插入散布圖



3.利用 *Excel* 求最佳直線：

⑴指令：*LINEST*

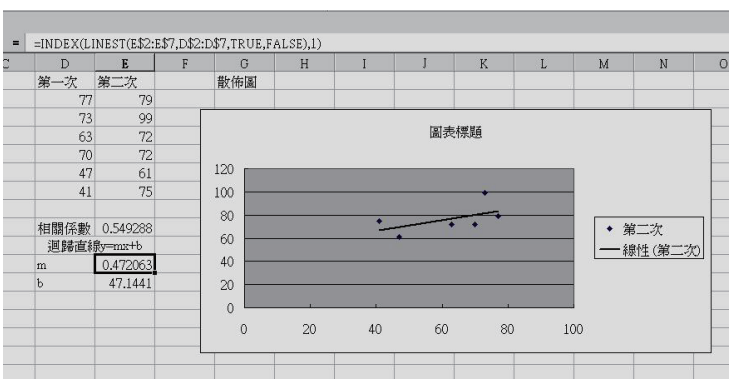
功用：使用最小平方法計算最適合於觀測資料組的迴歸直線公式，並傳回該直線公式的陣列。由於此函數傳回陣列值，所以必須輸入為陣列公式。

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

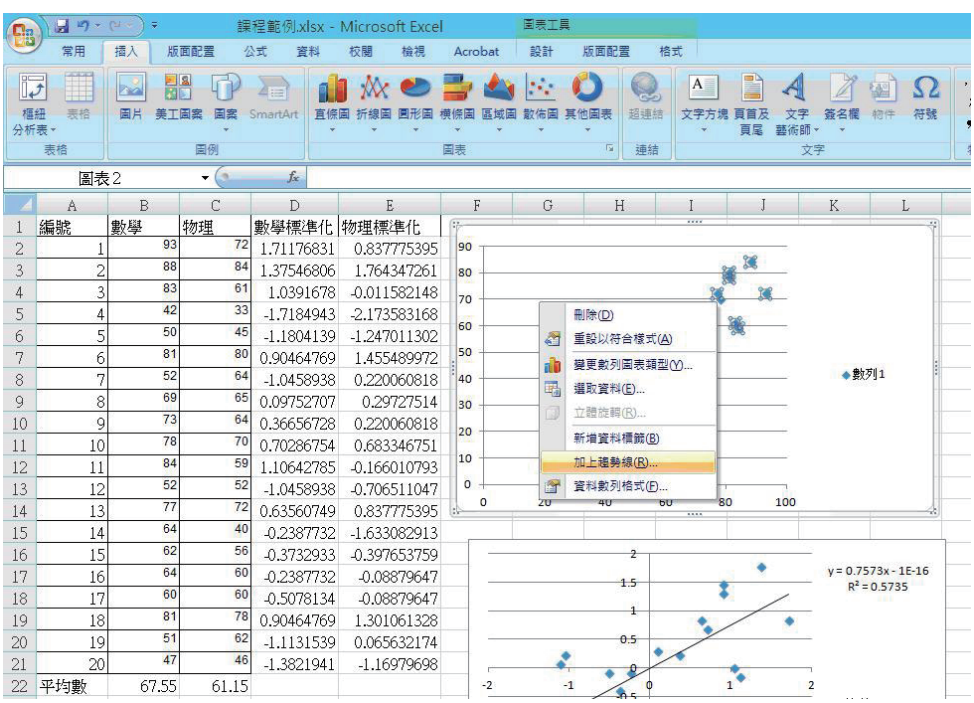
最佳直線：$y=mx+b$

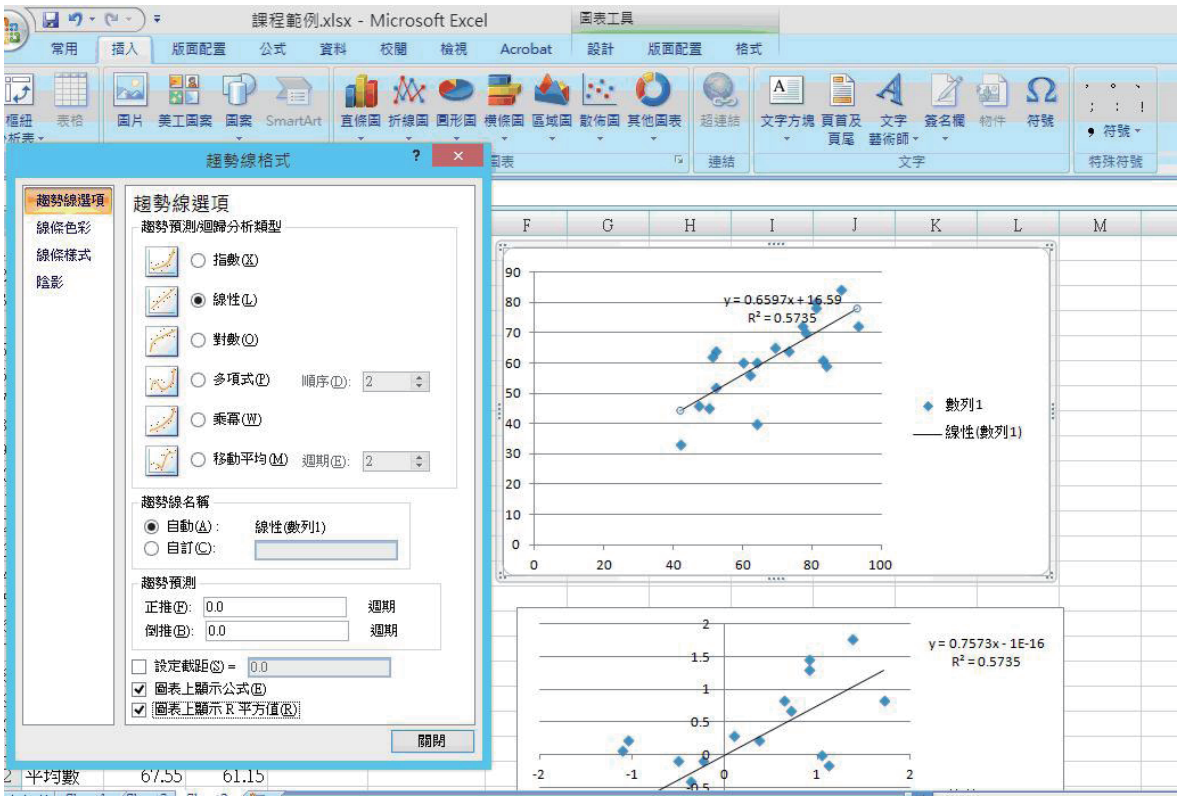
m 的計算：*INDEX（LINEST（known_y's,known_x's,const,stats）,1）*

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



⑵選定樣本點，然後按滑鼠右鍵再選加入趨勢線，再選單中選取線性，並且選圖表上顯示公式與 *R* 平方值 即可得到最佳直線與相關係數平方值。




課程範例.xlsx - Microsoft Excel
圖表工具
趨勢線格式
趨勢線選項
線條色彩
線條樣式
陰影
趨勢線選項
趨勢預測/迴歸分析類型
指數(X)
線性(L)
對數(O)
多項式(P)
順序(D):
乘冪(W)
移動平均(M)
週期(E):
趨勢線名稱
自動(A):
線性(數列1)
自訂(C):
趨勢預測
正推(F):
倒推(B):
週期
設定截距(S) =
圖表上顯示公式(E)
圖表上顯示 R 平方值(R)
關閉
y = 0.6597x + 16.59
R² = 0.5735
數列1
線性(數列1)
y = 0.7573x - 1E-16
R² = 0.5735
平均數
67.55
61.15

素養導向數學教材 / 曾世杰 主編
-- 初版 -- 新北市三峽區：國家教育研究院

1.數學教育
2.中小學教育
3.教材與教法

素養導向普通型高級中學數學教材：相關係數與最佳直線
主 編 者：單維彰
作　　者：林信安、曾俊雄
（依姓氏筆畫順序排列）
發 行 人：柯華葳
出 版 者：國家教育研究院
編 審 者：十二年國民基本教育數學素養教材研發編輯小組
召 集 人：曾世杰
副召集人：單維彰、鄭章華
編輯小組：古欣怡、朱安強、林美曲、林信安、馬雅筠、陳吳煜
陳淑娟、曾明德、曾俊雄、鄧家駿
（依姓氏筆畫順序排列）
版　　次：初版
電子全文可至國家教育研究院網站 *http: //www.naer.edu.tw* 免費取用

本書經雙向匿名審查通過