數據分析

1、一維數據分析（x_i表示$x_1, x_2, \cdots, x_n$）

(1) 集中趨勢數（具代表性）

(a) 平均（算術平均數）：$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n}$

（幾何平均數）：$G.M. = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$

(b) 眾數：M。表示出現次數最多的數。

(c) 中位數：M_e表示資料由小到大，最中間的數。

(2) 分散趨勢數（分散程度，數值恆正，且值越大表越 分散）

(a) 全距：R表示最大和最小數據的差。

(b) 標準差：$\sigma_x = \frac{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2}}{n} = \frac{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \mu_x^2}}{}$

（大約可想成與平均距離的平均）

(c) 變異數 σ_x^2。

2、標準分數 $\Rightarrow$ ① 平均 $= 0$ ② 標準差 $= 1$

設一組數據 $x_1, x_2, \cdots, x_n$ 的平均為 M，標準差為 σ，則

數據 x_i 的標準分數 $z_i = \frac{x_i - M}{\sigma_x}$。

註：將一組數據，減去其平均，再除以標準差，稱為標準化。

Ex1: 某項競賽為減少個人主觀影響，規定先將15位評審給同一位參賽者的成績求得算術平均數，再將與平均數相差超過1分的評審成績剔除後重新計算平均值做為此參賽成績。現在有位參賽者所獲15位評審的平均成績為76分，其中三位評審給92, 45, 55應剔除，則此參賽者的比賽成績為幾分？

Sol: ◎ 總分 $= \sum x_i = n \times$ 平均

15人總分 $= 76 \times 15 = 1140$

12人總分 $= 1140 - 92 - 45 - 55 = 948$

平均 $= \frac{948}{12} = 79$ #

Ex2: 小明觀察一組正整數，其中1有1個，2有2個，…，30有30個，則下列敘述何者正確？

(1) 此數據的眾數為30

(2) 此數據的中位數為20

(3) 此數據的算術平均是正整數

(4) 此數據的中位數大於算術平均數.

Sol: (1) 30出現最多次(0)

(2) 此數列共有 $\frac{30\times31}{2} = 465$ 個數

第233個數即為中位數

$\frac{n(n+1)}{2} \doteq 233 \Rightarrow \frac{20\times21}{2} = 210, \frac{21\times22}{2} = 232$

∴第232個數為21時的前一個 $\Rightarrow$ 中位數 $= 22$

(3)(4) 平均 $= \frac{1\times1+2\times2+\cdots+30\times30}{1+2+\cdots+30} = \frac{\frac{30\times31\times61}{6}}{\frac{30\times31}{2}}$

$= \frac{61}{3} <$ 中位數 $= 22$

Ex3: 小王沉迷於手機遊戲，於是媽媽規定小王必須降低手機的網路流量，平均每個月減少40%，已知規定的前三個月，網路流量分別較前一個月減少20%, 40%, 25%。則第四個月較前一個月減少 x%，小王才能達成規定。求 x。

Sol: 設起始流量 N (想剩下)

$N\cdot(0.6)^4 = N\cdot0.8\cdot0.6\cdot0.75\cdot(1-x\%)$

$\therefore N\cdot(\frac{3}{5})^3 = N\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot(1-x\%)$

$\therefore 1-x\% = \frac{9}{25} \therefore x\% = \frac{16}{25} = 64\%$ #

Ex4: 已知9筆數據，經標準化後之值分別為 $0.6, 0.7, 0.8, -0.1, -1.5, -2, 1, 0, x$。已知原來9筆數據的算術平均數為50，標準差為20，求原始數據的中位數。

Sol: ∵標準化後 $\Rightarrow$ 平均 $= 0$

$\therefore 0.6+0.7+0.8-0.1-1.5-2+1+0+x = 0$

$\Rightarrow x = 0.5$ ∴中位數 $= 0.5$

原始成績 $a \Rightarrow \frac{a-50}{20} = 0.5 \Rightarrow a = 60$ #

Ex5: 某生第一次段考考六科的平均為80分，若已知其中五科成績為 68, 80, 80, 80, 86，則其六科成績的標準差為何？

Sol:

六科總分 $= 80\times6 = 480$

∴第六科 $= 86$ 分

$\sigma = \sqrt{\frac{(-12)^2+0^2+0^2+0^2+6^2+6^2}{6}} = 6\sqrt{\frac{6}{6}} = 6$ #

Ex6. 某校高一第一次段考數學成績不佳，老師決定將每個人原始成績取平方根後再乘以10作為正式紀錄的成績。今抽100位同學發現調整後的成績平均為65分，標準差為15分；試問這100位同學未調整前的成績平均M介於哪一個連續整數間。

Sol:

原	→	新
x_i		$10\sqrt{x_i}$
平均		65
標準差		15

$$15=\sqrt{\frac{\sum(10\sqrt{x_i})^2}{100}-65^2}$$

$$\Rightarrow 225=\frac{100\sum x_i}{100}-4225$$

$$\therefore \sum x_i=4450 \Rightarrow \mu=\frac{\sum x_i}{100}=44.5 \#$$

Ex7. 陳老師將本學期任教果之兩班測驗成績結果如表所示。陳老師想將兩班共50人之成績合併，下列敘述何者正確？

(1) 合併之算術平均高於70分
(2) 合併之算術平均介於60至70之間
(3) 合併之中位數低於61分
(4) 合併之中位數介於61至68之間
(5) 合併之標準差低於8分

Sol: ◎ 平均 $\Leftrightarrow \frac{\sum x_i}{}$；標準差 $\Leftrightarrow \frac{\sum x_i^2}{}$

<性質> 合併後之平均、中位數必介於合併前間.

	甲	乙
平均	70	60
中位數	68	61
標準差	10	8
人數	30	20

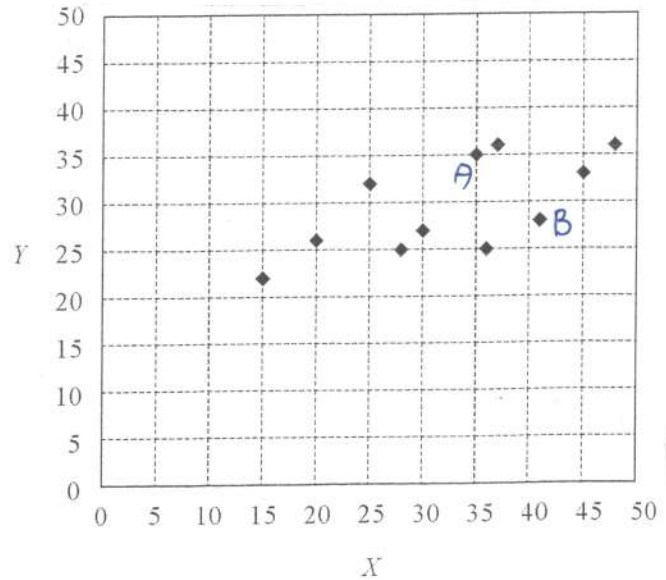
(1)(2) $\sum_{甲} x_i = 70\times 30 = 2100$
$\sum_{乙} x_i = 60\times 20 = 1200$
$\Big]\ \sum_{甲+乙} x_i = 3300$
$\mu_{合} = \frac{3300}{50} = 66$

(3)(4) 有 15+10 人低於 68
15+10 人高於 61

(5) $10=\sqrt{\frac{\sum_{甲} x_i^2}{30}-70^2} \Rightarrow \sum_{甲} x_i^2 = 150000$
$8=\sqrt{\frac{\sum_{乙} x_i^2}{20}-60^2} \Rightarrow \sum_{乙} x_i^2 = 73280$
$\Big]\ \sum_{合} x_i^2 = 223280$

Ex8. 某次測驗分成選擇題和非選擇題。下列的散佈圖中每個點(X,Y)分別代表學生此兩部分的得分，其中X表示選擇題的得分，Y表示非選擇題的得分。設 Z=X+Y 表示測驗總合，共有11位學生參加。試問下列選項何者正確？

(1) X中位數 > Y中位數
(2) X標準差 > Y標準差
(3) X全距 > Y全距
(4) Z中位數 = X中位數 + Y中位數



Ex9. 下列5組資料，何者標準差最大？

(1) A: 1, 1, 1, 1, 1, 10, 10, 10, 10, 10
(2) B: 1, 1, 1, 1, 1, 5, 5, 5, 5, 5
(3) C: 4, 4, 4, 5, 5, 5, 5, 6, 6, 6
(4) D: 1, 1, 2, 2, 3, 3, 4, 4, 5, 5
(5) E: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

$\sigma_{合}=\sqrt{\frac{223280}{50}-66^2}=\sqrt{109.6}$

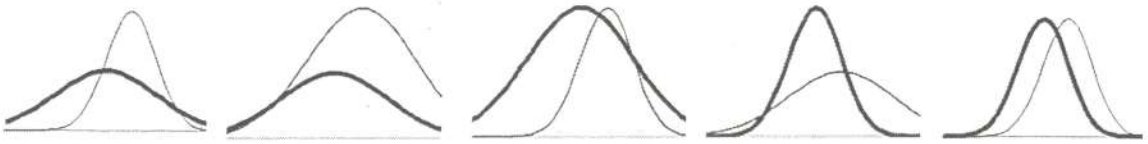
Sol: 資料A最分散 (1) #

Sol: 共11筆資料，中位數為第6筆資料

(1) X的中位數為A桌 ⇒ 35
Y的中位數為B桌 ⇒ <30
(2)(3) X落在15～50　X較Y分散
Y落在20～40
(4) ∵ X、Y之中位數是不同筆資料
∴ Z中位數 ≠ X中位數 + Y中位數

Ex10: 甲、乙兩校有一樣多的學生參加數學能力測驗，兩校學生測驗成績的分布都很接近常態分布，其中甲校學生的平均分數為 60 分，標準差為 10 分；乙校學生的平均分數為 65 分，標準差為 5 分。若用粗線表示甲校學生成績分布曲線；細線表示乙校學生成績分布曲線，則下列哪一個分布圖較為正確？（101 學測）

(1)　(2)　(3)　(4)　(5)



Sol:

	甲	乙
平均	60	65
標準	10	5

甲較分散且高峰較左邊

∵人數相同 ∴曲線下面積相同

(1)#

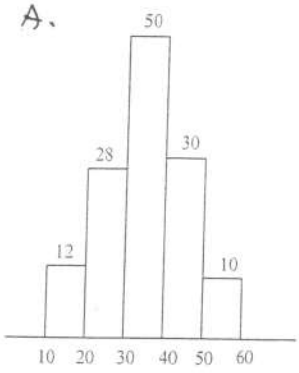
Ex11: 小明參加某次國文、英文、數學、自然、社會五個科目測驗，每一科分數均為 0～100 分。已知小明國英數三科分數分別為 75、80、85 分。試問下列哪些選項會讓五科成績的「平均不低於 80 分且標準差不大於 5 分」。

(1) 自然 75 分，社會 80 分
(2) 自然和社會皆為 80 分
(3) 自然和社會平均 85 分
(4) 自然和社會之和不低於 160 分且兩科差距不超過 10 分
(5) 自然和社會都介於 80～85 分之間

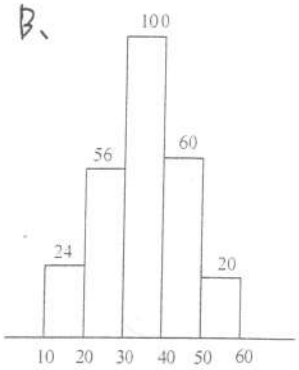
Sol:
(1) 平均未達 80
(2) 平均為 80 且與平均距離皆不超過 5
(3) 若自然 = 100，社會 = 70，與平均距離自然、社會均大於 10，國文大於 5
(4) 若自 = 社 = 100 分，與平均之距離有 4 個大於 5，1 個接近 5
(5) 平均略高於 80 分，與平均距離有 4 個小於 5，1 個接近 5

Ex12: 下列四個直方圖，其標準差分別為 σ_A、σ_B、σ_C、σ_D，試比較其大小。

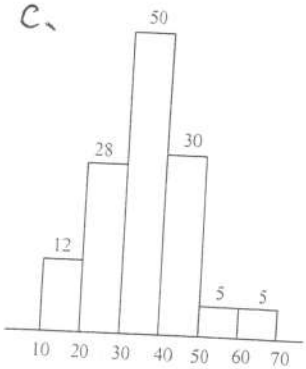
A、



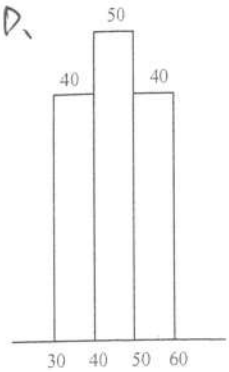
B、



C、



D、



Sol: $\sigma_C > \sigma_A = \sigma_B > \sigma_D$

3、二維數據分析 (x_i,y_i) 表 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$

(1) 相關係數：$r_{xy}=\dfrac{\Sigma(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\Sigma(x_i-\mu_x)^2}\sqrt{\Sigma(y_i-\mu_y)^2}}=\dfrac{\Sigma x_iy_i-n\mu_x\mu_y}{\sqrt{\Sigma x_i^2-n\mu_x^2}\sqrt{\Sigma y_i^2-n\mu_y^2}}$。

(a) 意義：將 (x_i,y_i) 描點在散佈圖上的圖形多像 直線。

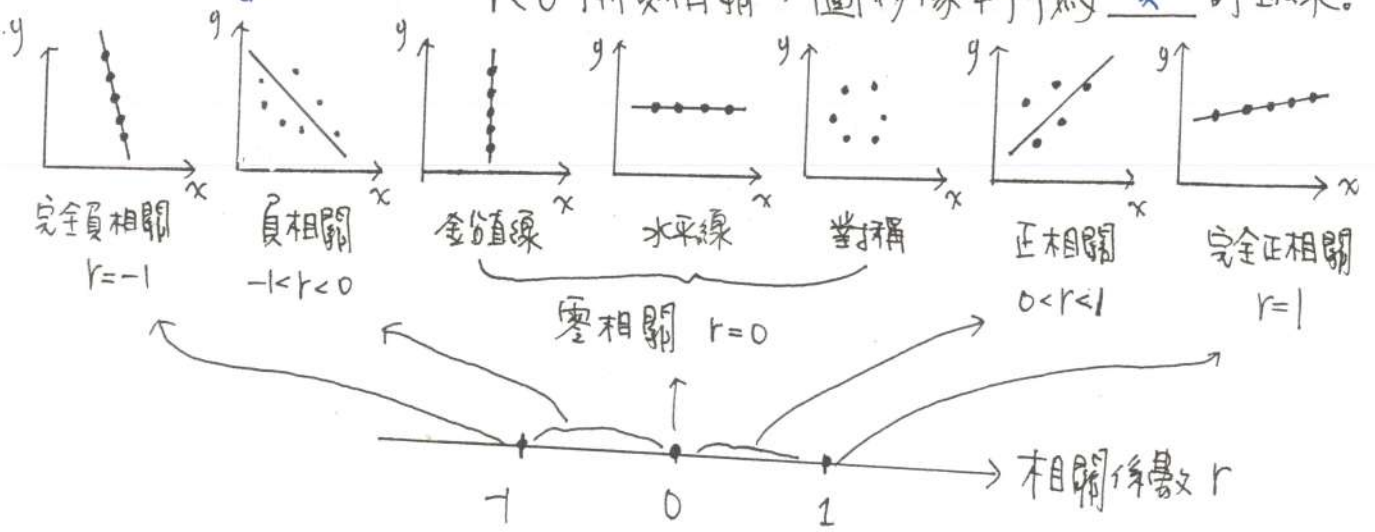
(b) 性質：① r 的範圍 ⇒ $-1\le r\le 1$。（鉛直線和水平線除外）

① 數值 $|r|$：$|r|$ 越大，表示 x,y 相關程度大 ⇒ 圖形越像直線。

正負：$r>0$ 稱正相關 ⇒ 圖形像斜率為 正 的直線。

$r<0$ 稱負相關 ⇒ 圖形像斜率為 負 的直線。

更正

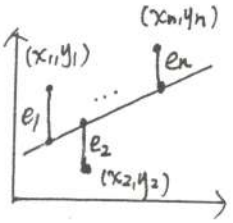


(2) 迴歸線：散佈圖中，很像的那條直線，稱為迴歸線。

(a) 最小平方法：設 迴歸線方程式 $Y=ax+b$，必滿足

殘差平方和 $e_1^2+e_2^2+\cdots+e_n^2$ 最小。

（即 $(ax_1+b-y_1)^2+(ax_2+b-y_2)^2+\cdots+(ax_n+b-y_n)^2$）



(b) 迴歸線方程式：$y-\mu_y=r\cdot\dfrac{\sigma_y}{\sigma_x}(x-\mu_x)$（① $m=r\cdot\dfrac{\sigma_y}{\sigma_x}$ ② 過 (μ_x,μ_y)）

若將資料 (x_i,y_i) 標準化得新數據 (x_i',y_i')，

標準化數據 (x_i',y_i') 的 迴歸線方程式為 $y'=rx'$。

[轉化]

$$y'=rx'\Rightarrow\frac{y-\mu_y}{\sigma_y}=r\cdot\frac{x-\mu_x}{\sigma_x}\Rightarrow y-\mu_y=r\cdot\frac{\sigma_y}{\sigma_x}\cdot(x-\mu_x)$$

Ex13: 某肥皂廠推出新產品，在上市前以不同的單價 x（單位：十元）調查市場的需求量 y（單位：萬盒）。調查結果如下表，求(1) x 和 y 的相關係數
(2) y 對 x 的迴歸線方程式

x	8	9	10	11	12	$\Rightarrow \mu_x=10$
y	11	12	10	8	9	$\Rightarrow \mu_y=10$

Sol:

$x-\mu_x$	-2	-1	0	1	2
$y-\mu_y$	1	2	0	-2	-1

(1)
$$r=(\cos\theta)=\frac{-2-2+0-2-2}{\sqrt{4+1+1+4}\sqrt{1+4+4+1}}=\frac{-4}{5}$$

(2) $\sigma_x=\sqrt{\frac{10}{5}}$, $\sigma_y=\sqrt{\frac{10}{5}}$

$$\therefore y-10=\frac{-4}{5}\cdot\frac{\sqrt{2}}{\sqrt{2}}(x-10)$$

$$\therefore y-10=\frac{-4}{5}(x-10) \quad \#$$

Ex14: 調查某國家一年5個地區，香菸與肺癌之相關性，所得到的數據為 (x_i,y_i)，$i=1,2,3,4,5$，其中變數 X 表示每人香菸消費量（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：

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

$\sum_{i=1}^{5} y_i=105$, $\sum_{i=1}^{5} y_i^2=2209$, 求

(1) X 和 Y 的相關係數

(2) 若甲每年香菸消費 67（單位：十包），乙每年香菸消費 35（單位：十包），則依迴歸直線預測甲罹患肺癌是乙罹患肺癌的幾倍？

Sol: (1)
$$r=\frac{\sum x_iy_i-n\mu_x\mu_y}{\sqrt{\sum x_i^2-n\mu_x^2}\sqrt{\sum y_i^2-n\mu_y^2}}=\frac{7}{12}\ \#$$

$\left(\mu_x=\frac{\sum x_i}{5}=27,\ \mu_y=\frac{\sum y_i}{5}=21\right)$

$$=\frac{2842-5\times27\times21}{\sqrt{3661-5\times27^2}\sqrt{2209-5\times21^2}}=\frac{7}{4\times3}$$

(2) $\sigma_x=\sqrt{\frac{3661-5\times27^2}{5}}=\frac{4}{\sqrt{5}}$, $\sigma_y=\sqrt{\frac{2209-5\times21^2}{5}}=\frac{3}{\sqrt{5}}$

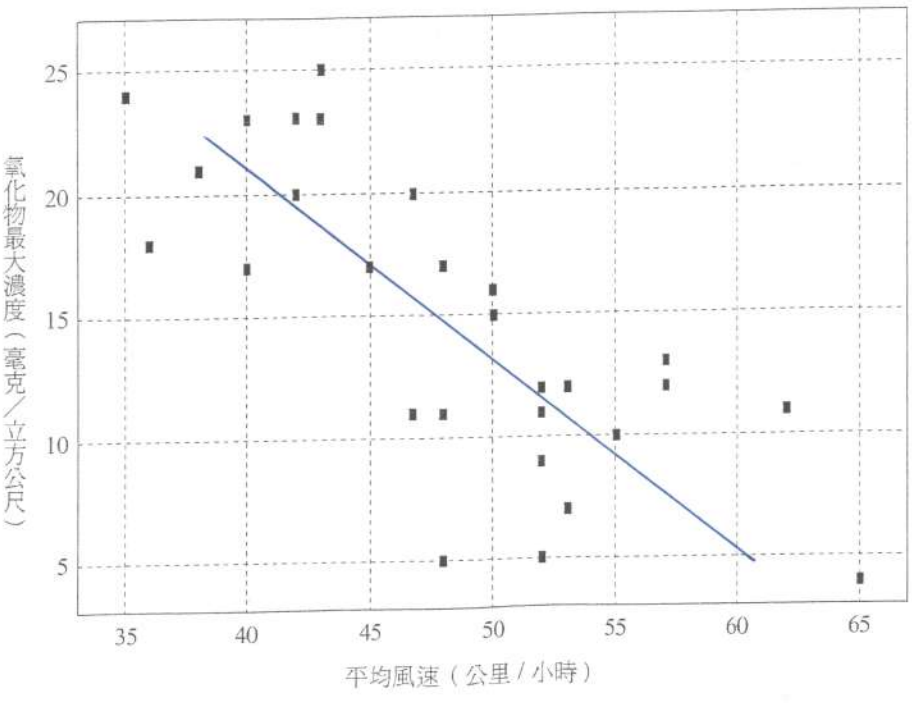
$$\therefore y-21=\frac{7}{12}\times\frac{\frac{3}{\sqrt{5}}}{\frac{4}{\sqrt{5}}}(x-27)\Rightarrow y=\frac{7}{16}(x-27)+21$$

甲：$y=\frac{7}{16}\times40+21=\frac{77}{2}$

乙：$y=\frac{7}{16}\times8+21=\frac{49}{2}$

$\therefore \frac{11}{7}$ 倍 #

Ex15: 某機構為瞭解特定區域的空氣品質，連續二十八天蒐集了該地區的平均風速及特定氧化物的最大濃度，再繪製散佈圖(如下)，根據該圖可知此資料中



(1) 該氧化物最大濃度的標準差大於15
(2) 該氧化物最大濃度的中位數為15
(3) 平均風速的中位數介於45～50
(4) 若以最小平方法決定數據集中趨勢的直線，則該線斜率小於0

Sol: (1) 落在 5～25，平均約 15 $\Rightarrow \sigma<15$
(2) 中位數為第14和15筆平均數=15
(3) 45左邊10筆，50右邊11筆 P6、

Ex16: 小明參加某次路跑 10 公里組的比賽，下表為小明手錶所記錄之各公里的完成時間、平均心率及步數：

	完成時間	平均心率	步數
第一公里	5:00	161	990
第二公里	4:50	162	1000
第三公里	4:50	165	1005
第四公里	4:55	162	995
第五公里	4:40	171	1015
第六公里	4:41	170	1005
第七公里	4:35	173	1050
第八公里	4:35	181	1050
第九公里	4:40	171	1050
第十公里	4:34	188	1100

在這 10 公里的比賽過程，請依據上述數據，選出正確的選項。

(1) 由每公里平均心率得知小明最高心率為188

(2) 每公里完成時間和平均心率為正相關

(3) 每公里完成時間和步數為負相關

(4) 小明此次路跑，每步距離平均小於1公尺

(5) 每公里步數和平均心率為正相關

(3) 整體而言，時間越多，心率越低⇒負

(5) 〃，步數越多，心率越高⇒正

(5) 〃，時間越多，步數越少⇒負

Sol: (1) 平均心率最高188 ⇒ 最高心率 ≥ 188

(2) 每公里平均步數 > 1000

⇒ 每步距離平均 < $\frac{1\text{公里}}{1000}$ = 1(m)

Ex17: 下圖是國文、英文、歷史三科成績分布直方圖，求下列哪些推論正確：

(1) 歷史平均分數比國文平均分數低

(2) 歷史平均分數最低

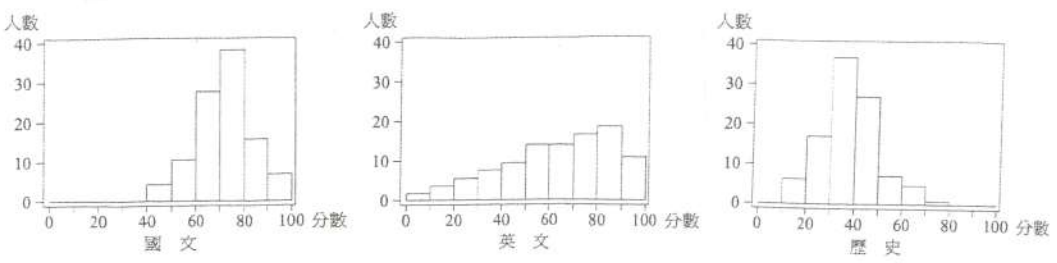
(3) 英文標準差比國文標準差小(大)

(4) 英文標準差最大

(5) 「國文與歷史之相關係數」比「國文與英文之相關係數」高.

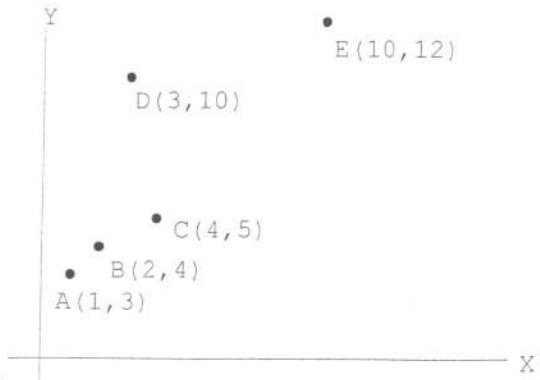
Sol: (2)(4) 英文成績較分散

(5) 無法判定國文高，是否歷史高or英文高



Ex18: 如圖所示有5筆(X,Y)資料。試問去掉哪一筆資料後，剩下4筆資料對相關係數最大？

Sol: D#



P7.

4. 資料的平移、伸縮：

	集中趨勢數			分散趨勢數		相關係數
	平均	眾數	中位數	全距	標準差	設 X, Y 相關係數 r_{XY}
X	M_x	M_0	M_e	L	σ_x	$X'=aX+b$; $Y'=cY+d$
$X'=aX+b$	aM_x+b	aM_0+b	aM_e+b	$\lvert a\rvert \cdot L$	$\lvert a\rvert \cdot \sigma_x$	$\Rightarrow r_{X'Y'}=\frac{ac}{\lvert ac\rvert} r_{XY}$

加、減、乘、除跟著變　　加、減不變，乘、除跟著變　　加、減、乘、除均不影響，僅正、負影響。

Ex19： 甲、乙、丙三人參加模測，成績如下表，S_1, S_2, S_3 分別代表甲、乙、丙三人五科的標準差，求 S_1, S_2, S_3 的大小關係。

	社會	國文	自然	英文	數學
甲	100	70	80	60	50
乙	90	60	70	50	40
丙	80	56	64	48	40

Sol：

乙 = 甲 − 10 $\Rightarrow S_2 = S_1$

丙 = 甲 × 0.8 $\Rightarrow S_3 = 0.8 S_1$

$\therefore S_1 = S_2 > S_3$ #

Ex20： 根據統計，1月台北平均溫度攝氏16度，標準差攝氏3.5度。已知攝氏 x 度時，華氏溫度為 $y=\frac{9}{5}x+32$。若用華氏溫度表示，則台北平均溫度＿度，標準差＿度。

Sol：

$M_y = \frac{9}{5} M_x + 32 = \frac{9}{5} \times 16 + 32 = 60.8$ #

$\sigma_y = \frac{9}{5} \sigma_x = \frac{9}{5} \times 3.5 = 6.3$ #

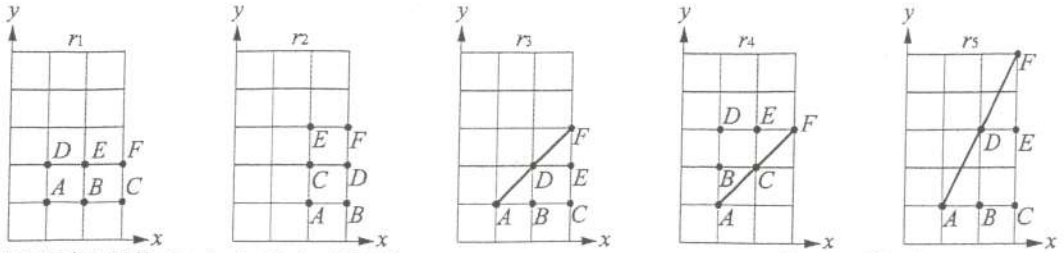
Ex21： 令 X 代表每個高中生平均每天研讀數學的時間(以小時計)，則 $W=7(24-X)$ 代表每個高中生平均每週花在研讀數學以外的時間。令 Y 代表高中生數學學測成績。設 X, Y 相關係數 R_{XY}，W, Y 相關係數 R_{WY}，則 R_{XY} 和 R_{WY} 之關係，何者正確？

(1) $R_{WY}=7(24-R_{XY})$　(2) $R_{WY}=7R_{XY}$　(3) $R_{WY}=-7R_{XY}$　(4) $R_{WY}=R_{XY}$　(5) $R_{WY}=-R_{XY}$

Sol： $W=-7X+168$

$\therefore R_{WY}=-R_{XY}$　(5) #

Ex22: 下圖中，有五組數據，每組各有 A、B、C、D、E、F 等六個資料點。(86 推甄)



Sol: 設各組的相關係數由左至右分別為 r_1, r_2, r_3, r_4, r_5，試比較其大小關係。

$r_1 = r_2 = 0$

$r_3 = r_4 = r_5 > 0$

(三) → (四)

$(x, y) \to (y, x)$

(三) → (五)

$(x, y) \to (x, 2y-1)$

Ex23: 高三 300 位學生，數學科以 X、Y 分別表示第一次、第二次段考成績。若這兩次段考的相關係數為 0.016，則下列那些是正確的？

(1) ✓ X、Y 的相關情形可以用散布圖表示

(2) 這兩次段考成績適合用直線 $Y = a + bX$ 表示相關情形

(3) $X+5$ 和 $Y+5$ 的相關係數為 0.016

(4) ✓ $10X$ 和 $10Y$ 的相關係數為 0.016

(5) $X' = \frac{X - M_X}{\sigma_X}$，$Y' = \frac{Y - M_Y}{\sigma_Y}$，其中 M_X, M_Y 分別為 X、Y 平均；σ_X, σ_Y 分別為 X、Y 標準差，則 X'、Y' 的相關係數為 0.016

Sol: (1) 相關情形皆可用散布圖表示

(2) $r = 0.016$ 接近 0 ⇒ 不適合

(3) 加、減、乘、除皆不改變，僅受正負影響

Ex24: 英國某實驗研究一金屬圓柱在不同負重下對柱高的影響。其結果如下：(0, 70.5), (2, 69.4), (4, 68.4), (6, 67.2), (8, 66.3), (10, 65.5), (12, 64.4)

其中測量單位分別為英吋和英寸。將此筆資料的相關係數記為 r，以最小平方法決定的直線斜率記為 m。將單位換為公吋與公分（1 英吋 = 1.016 公吋，1 英寸 = 2.54 公分），若單位轉換後的相關係數記為 R，以最小平方法決定的直線斜率記為 M。下列何者正確？

(1) ✓ $r \cdot m > 0$ (2) $r > 0$ (3) ✓ $r = R$ (4) $m = M$

Sol: $(x, y) \to (x', y')$

$x' = 1.016x \Rightarrow R = r$

$y' = 2.54y$

∵ x 越大，y 越小 ⇒ $r < 0$

(1) $r \cdot m = r \cdot r \cdot \frac{\sigma_y}{\sigma_x} = r^2 \frac{\sigma_y}{\sigma_x} > 0$

(2) $r < 0$

(3) $r = R$

(4) $M = R \cdot \frac{\sigma_{y'}}{\sigma_{x'}} = r \cdot \frac{2.54\sigma_y}{1.016\sigma_x} = \frac{2.54}{1.016} m$