統計學

下冊

鄭堯栚著

商務印書館發行

統計學

下冊

鄭堯栚著

商務印書館發行

中華民國二十九年五月初版

㊣統計學二冊

㊣(3 2 0 7 4)

每部實價國幣伍元
外埠酌加運費匯費

著作者 鄭堯枏

發行人 王雲五 長沙南正路

印刷所 商務印書館

發行所 商務印書館 各埠

（本書校對者施伯朱胡達聰）

第四編　相關論

第一章　簡單相關

第一節　總述

依前述知在含有一種事象之統計數列上，得依平均數差量及偏斜度等以表示其性質。但在並列二種或二種以上事象之統計數列時，則其事象間之相互關係亦有研究之必要。例如夫妻之年齡間，在丈夫之年齡高大者其妻之年齡一般亦隨之高大，逆亦成立，知二者間有正的相關。反之在農產物之收穫量與其價值間，若不考慮其他各種情形，則知收穫量衆多時其價值趨於低廉，逆亦成立，知二者間有逆的相關。此相關非只學問上之研究而已，且因吾人之世界是爲相關之世界，常識之大部分其相互間均有關係之存在，只其程度有深淺之不同，故知對於測定相關程度一事，非只經濟學，生物學，心理學，教育學等之研究上所必需，在吾人之日常生活上亦甚重要。以後將測定相關程度之方法，依次說明：

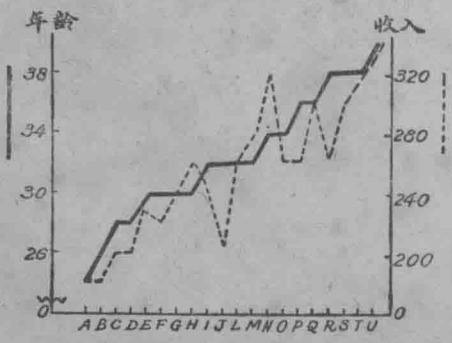
第二節　相關表

由含有二種事象之統計數列所求得之相關，稱爲簡單相關。但欲測定此相關程度時，則有作成相關表 (correlation table) 之必要。例如測

量某公司店員之收入與年齡間之關係起見，作成有如第 113 表之統計表，現由此資料將人名取在橫軸上，年齡取在左方縱軸上，收入取在右方縱軸上以畫得一圖形如第 123 圖，依此圖形在大體上得知年長者之收入比年小者爲多，其年齡與收入間存有一種相關。若將第 113 表內情形改書爲第 114 表之格式以表現時，知表內第一縱行與第一橫行相交

第 113 表

人名	年齡	收入	人名	年齡	收入
A	24	180	L	32	260
B	26	180	M	32	280
C	28	200	N	34	320
D	28	200	O	34	260
E	30	240	P	36	260
F	30	220	Q	36	300
G	30	240	R	38	260
H	30	260	S	38	300
I	32	240	T	38	320
J	32	200	U	40	340



第 124 圖

表 114 第　收入（單位十元）

年齡（歲）	18	20	22	24	26	28	30	32	34	計
24	1									1
26	1									1
28		2								2
30			1	2	1					4
32		1		1	1	1				4
34					1			1		2
36					1		1			2
38					1		1	1		3
40									1	1
計	2	3	1	3	5	1	2	2	1	20

處之數字 1,是表示收入 180 元年齡 24 歲者一人（即 A),第二縱行與第三橫行相交處之數字 2,是表示收入 200 元年齡 28 歲者二人(即 C 與 D)，餘類推。其最右端寫有計字之縱行上各數字,是各相當橫行內所有各數字相加而成。例如 30 歲所當橫行內最右端數字 4,是由該行內所有數字 1,2,1 相加而成，由此即知年齡 30 歲者共有四人。同樣在表之左端最下處寫有計字之橫行內各數字，是由所當各縱行上所有數字相加而成。例如第五縱行（收入爲 260 元）最下端之數字 5 是由該縱行內所有數字 1,1,1,1,1 等相加而成，由此知年收入 260 元者共五人，餘類推。最後其最右之數字 20，爲最右計之縱行內各數之和,亦爲最下橫行內各數字之和,即

$$1+1+2+4+4+2+2+3+1=20,$$

$$2+3+1+3+5+1+2+2+1=20,$$

此 20 人爲公司店員之總人數。若雙方各自相加和數不能相等時，則其二者中必有一方錯誤，須卽行檢查以訂正之。

如此所作成表，稱爲該公司店員年齡與收入間之相關表。本表是將 20 名店員依其年齡與收入詳細分類，使二者間之相關明白表現，知其年齡大(小)者收入爲多(少)。但此等情形在第 113 表上，實難以尋得。在次數稀少之例上相關表已有如此之優越，何況在次數衆多之資料上，其使用相關表之便利不言可知矣。

現爲增進議論之便利起見，特將相關表依一般記號說明如次：

設二變量 X, Y 之分類組別爲

$$X_1,\ X_2,\cdots\cdots X_m\ (m\ 個),$$

$$Y_1,\ Y_2,\cdots\cdots Y_n\ (n\ 個),$$

其 X_i 與 Y_k 所組合次數用 f_{ik} 以表示時，得其相關表如第 115 表。

第 115 表

Y \ X	X_1	X_2	⋯	X_i	⋯	X_m	f	M
Y_1	$f_{1,1}$	$f_{2,1}$	⋯	$f_{i,1}$	⋯	$f_{m,1}$	f_1	M_1
Y_2	$f_{1,2}$	$f_{2,2}$	⋯	$f_{i,2}$	⋯	$f_{m,2}$	f_2	M_2
⋮	⋮	⋮	⋯	⋮	⋯	⋮		⋮
Y_k	$f_{1,k}$	$f_{2,k}$	⋯	$f_{i,k}$	⋯	$f_{m,k}$	f_k	M_k
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮	⋮
Y_n	$f_{1,n}$	$f_{2,n}$	⋯	$f_{i,n}$	⋯	$f_{m,n}$	f_n	M_n
f'	f'_1	f'_2	⋯	f'_i	⋯	f'_m	N	M_x
M′	M'_1	M'_2	⋯	M'_i	⋯	M'_m	M_c	

表內之

$$f_k = f_{1,k} + f_{2,k} + \cdots\cdots + f_{m,k},$$

$$f_i' = f_{i,1} + f_{i,2} + \cdots\cdots + f_{i,n},$$

$$N = \sum_k f_k = \sum_i f_{io}$$

在相關表上普通有求出各縱行平均與各橫行平均之必要。例如欲研究女子之多產性是否遺傳時，須先知生產小孩一人之母親的女兒平均生產幾人，生產小孩二人之母親的女兒平均生產幾人……等，而後始能研究母親之生產數均多者其女兒之生產數在平均上是否亦隨之衆多，或女兒生產數均多者其母親之生產數在平均上是否衆多等問題。故知相關表內有時須計算縱橫各行之平均，其平均數之求法如次：

$$M_k = \frac{\sum_i f_{i,k}\, Y_i}{\sum_i f_{i,k}} = \frac{1}{f_k} \sum_{i=1}^{m} f_{i,k}\, X_i,$$

$$M_i' = \frac{\sum_k f_{i,k}\, Y_k}{\sum_k f_{i,k}} = \frac{1}{f_i'} \sum_{k=1}^{n} f_{i,k}\, Y_k. \qquad \text{(公式 85)}$$

至變量 X, Y 之在全體上的平均數求法如次：

$$M_X = \frac{\sum_k f_k M_k}{\sum_k f_k} = \frac{\sum_i f_i' X_i}{\sum_i f_i'}$$

$$M_Y = \frac{\sum_i f_i' M_i'}{\sum_i f_i'} = \frac{\sum_k f_k Y_k}{\sum f_k} \qquad \text{(公式 86)}$$

此 M_X 稱爲 X 之總平均數，M_Y 稱爲 Y 之總平均數。

在相關表內其次數

(1)集中在其由左上角至右下角之對角線附近者，成爲正相關(direct correlation)。

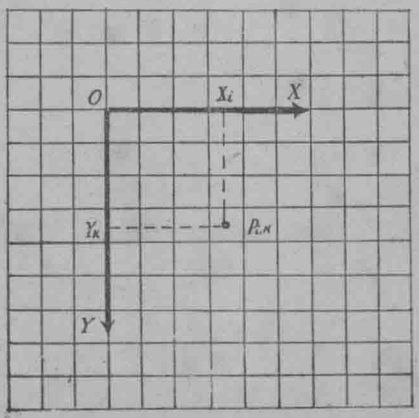
(2)集中在其由左下角至右上角之對角線附近者成爲負相關(inverse correlation)。

(3)不集中於此等對角線之附近者，其相關甚微。

注意：相關表內 Y 變數之排列，爲與次數分配表相一致起見，仍取上小下大。但在許多統計書內是上大下小，對此宜注意之。

第三節 相關圖形

當研究相關時，除前節準備外，尚須有圖學上之準備。在方格子紙上任意取一點 0，過此 0 點之橫軸上取 $X_1, X_2, \cdots\cdots X_m$，縱軸上取 $Y_1, Y_2, \cdots\cdots Y_n$ 等值，其橫線 X_i 與縱線 Y_k 相交點看做爲 $f_{i,k}$ 點相疊積處，其 0 點勿必一定取在座標軸 (X, Y) 之原點(X=0, Y=0) 處，依變量之範圍適宜決定之可也。至座標之方向爲使與相關表一致起見，特取 Y 軸之方向爲由上向下，與普通之座標方向相反，至 X 軸



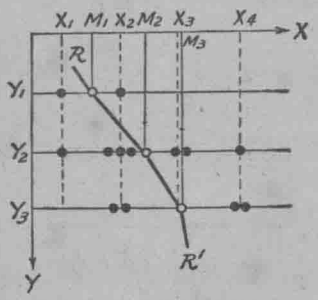
第 125 圖

之方向則仍舊。又爲說明便利起見，以後對於 X_i 與 Y_k 之相交點，略稱之爲點(X_i, Y_k)，其情形如第 124 圖。

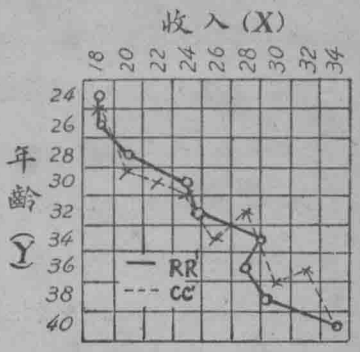
次在 X 軸上取各平均數 M_k 所當點，Y 軸上取各平均數 M_i' 所當點，連接 (M_1, Y_1)，(M_2, Y_2)……(M_n, Y_n) 等點得一折線 RR′（如第 125 圖）此折線 RR′ 稱爲橫行之迴歸線或稱爲對於 Y 之 X 的迴歸線 (line of regression of X on Y)。

同樣，連接 (X_1, M_1')，(X_2, M_2') …… (X_m, M_m') 等點得一折線 CC′ 此折線 CC′ 稱爲縱的迴歸線或稱爲對於 X 之 Y 的迴歸線 (line of regression of Y on X)。

在一相關表內定能作成二迴歸線，例如依第 114 表資料作成 RR′ 與 CC′ 二迴歸線如第 126 圖如此所得圖形稱爲相關圖形。



第 126 圖



第 127 圖

在 RR′ 上任意取一點 (M_k, Y_k)，此點爲表示第 k 橫行上 (Y_k) 所有 f_k 個單位值 X 之平均點，又依平均數之性質，M_k 是爲 Y_k 所

當 f_k 個單位值 X 之代表物，故亦得看做爲 Y_k 所當 f_k 個次數均疊積在此點 (M_k, Y_k) 上，依此知 Y 值各爲 $Y_1, Y_2, \cdots\cdots Y_n$ 時，其所對應之 X 值在上述理論之下，得看做爲 $M_1, M_2, \cdots\cdots M_n$。但 RR′ 是連接此等點而成，依此知 RR′ 實得看做爲表示種種 Y 值所對應之 X 值的曲線；換言之，迴歸線 RR′ 是將 Y 值所對應之 X 值依統計的(平均的)方法以說明其間之變動情形。

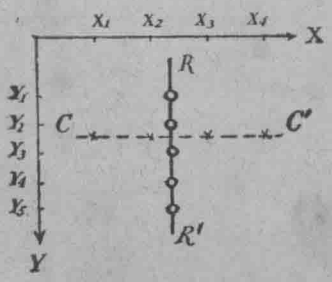
同樣，知縱行迴歸線 CC′，是將 X 值所對應之 Y 值依統計的(平均的)方法以說明一切。

迴轉線之名稱發源於 F. Galton (1822-1907)，彼在研究父子身長間之相關時，發見在平均上子方由其平均身長而來之偏差常小於父方由平均身長而來之偏差，依此知子方身長常有迴歸於其平均身長之趨勢，於是將表示此等關係之平均身長線，稱之爲迴歸線。

上述二迴歸線在相關理論上占重大任務，現爲便於進一步研究起見，特先就其極端者與以一種考慮。

(1)相關完全不成立者。

當此時其相關圖形成爲第 127 圖之狀態，其迴歸線變爲與二軸相平行之直線，其二變數 X, Y 中，雖任意變動其一方數值，仍不致發生影響於其所對應之他方的平均數上，知 X 與 Y 間成爲完全無關。依此知在作成相關圖形時，若其二迴歸線與二軸相平行時，卽知該二事象間完全無相關

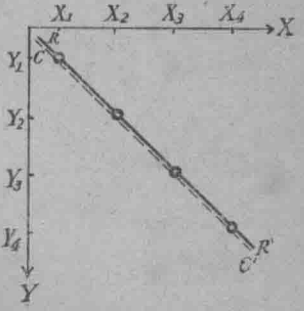


第 128 圖

之存在。

(2)相關完全存在者

在物理化學等精密科學上，依二事象 X,Y 間所存在之法則，能使一方給與數值時，他方數值卽能由此以決定之，其 X,Y 間成爲一種函數的關係。有如此性質之 X,Y 間實成爲完全之相關，而其所作成之二迴歸線完全互相一致，如第 128 圖。



第　129　圖

但在實際之統計問題上，能完全無相關及有完全相關者均甚稀少，故吾人研究之目的，只要放在實際上以研究其依如何方法始能精密測定相關之程度已足。對此研究上，其先決問題，是爲如何去決定依 Y(X) 之變化所起 X(Y) 上各變化之近似的法則。對此上述之迴歸線 RR′ (CC′) 實能滿足之。蓋因 $M_1, M_2, \cdots\cdots M_n$ ($M_1', M_2', \cdots\cdots M_m'$) 等爲 $X_1, X_2 \cdots\cdots$ ($Y_1, Y_2, \cdots\cdots$)等之代表物，故只要明白 Y_k (X_1) 與 M_k (M_i') 間之關係，卽能測得 Y (X) 與 X (Y) 間之關係，由此知迴歸線 RR′ (CC′) 實能作爲依 Y (X) 之變化所起對應值 X (Y) 上各變化之近似的法則。不幸迴歸線普通是爲折線，其表示既已複雜，而法則之計算又甚困難，以致難達研究之目的，依此不得不別想方法，以期能求得最簡之近似法則，但依解析幾何學，知研究 X,Y 間關係之最簡者，是爲直線，其方程式爲

$$Y = AX + B, \qquad (A,B = \text{常數})。$$

因此特將 Y (X) 之變化所起 X (Y) 上各變化之近似的法則，由迴歸線改換爲迴歸直線 (stright line of regression)。依此改換後。吾人只要研究一次方程式卽能明瞭其近似的法則之一般情形。對此變換上，在迴歸線上含有各特質，雖不免有所相差，但在大體上已足吾人應用。

第四節 迴歸直線

於相關圖形上取第二橫行 Y_2 上各點如次：

點之名稱： $P_{1,2}$, $P_{2,2}$, $P_{3,2}$, $\cdots\cdots P_{m,2}$,

X 之數值： X_1, X_2, X_3, $\cdots\cdots X_m$,

相疊點數： $f_{1,2}$, $f_{2,2}$, $f_{3,2}$, $\cdots\cdots f_{m,2}$,

在此橫行上各 X 之算術平均數爲 M_2，此 M_2 所當點是將 Y_2 行上所有各分布值 (X) 與此點相差之偏差的標準差成爲最小之點(最小偏差點)，故知

$$\sqrt{\frac{f_{1,2}\overline{P_{1,2}A}^{2}+f_{2,2}\overline{P_{2,2}A}^{2}+\cdots\cdots+f_{m,2}\overline{P_{m,2}A}^{2}}{f_{1,2}+f_{2,2}+\cdots\cdots+f_{m,2}}}$$

$$=\sqrt{\frac{\sum_{i} f_{i,2}(X_i-A)^2}{f_2}}$$

成爲最小時，其式內之 A 值實爲 M_2。

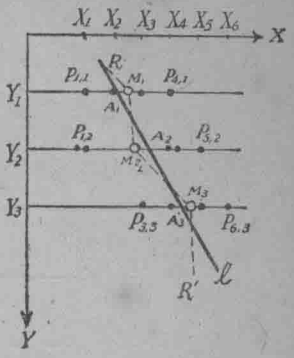
依此知迴歸線 RR′ 是爲將橫行 $Y_1, Y_2, \cdots\cdots Y_n$ 上所有各點 X 之標準差變爲最小各點 $M_1, M_2, \cdots\cdots M_n$ 等所接成之折線，並能給與

由 Y 值之變化所對應諸 X 值所起變化之近似的法則，其圖示情形如第 129 圖。

但吾人要求之目的，依前節知爲欲改此折線爲直線以表示其近似的法則，故對於變換上須滿足次之二項條件：

(1)須變爲直線。

(2)須與迴歸線之作成方法相似以進行一切。



第 130 圖

欲達到此等目的時，必須先由一直線 l 至相關圖形上各點 $P_{11}, P_{21}, \cdots\cdots; P_{12}, P_{22}, \cdots\cdots; P_{13}, P_{23}\cdots\cdots$ 等所生偏差之標準差上着手進行而後始可。

在第 129 圖內橫行 $Y_1, Y_2, \cdots\cdots Y_n$ 與直線 l 相交點各爲 $A_1, A_2, \cdots\cdots A_n$，在其第一橫行 Y_1 上所有各點與 A_1 之偏差爲

$$\overline{P_{11}A_1}, \overline{P_{21}A_1}, \cdots\cdots \overline{P_{i1}A_1}, \cdots\cdots \overline{P_{m1}A_1}$$

第二橫行 Y_2 上所有各點與 A_2 之偏差爲

$$\overline{P_{12}A_2}, \overline{P_{22}A_2}, \cdots\cdots \overline{P_{i2}A_2}, \cdots\cdots \overline{P_{m2}A_2}$$

$$\cdots\cdots$$

第 n 橫行 Y_n 上所有各點與 A_n 之偏差爲

$$\overline{P_{1n}A_n}, \overline{P_{2n}A_n}, \cdots\cdots \overline{P_{in}A_n}, \cdots\cdots \overline{P_{mn}A_n}$$

時，在此等偏差 $(P_{ik}A_k = X_i''_k)$ 之自乘上各乘以其所當之重疊點數

$$f_{1,1}, f_{2,1}, \cdots\cdots f_{i,1}, \cdots\cdots f_{m,1}$$

$$f_{1,2}, f_{2,2}; \cdots\cdots f_{i,2}, \cdots\cdots f_{m,2}$$

……

$f_{1,n}$, $f_{2,n}$, …… $f_{i,n}$ …… $f_{m,n}$

後舉行相加，再於其結果上除以平面上所有各點之總數後並再施以開方，卽得次量

$$\sqrt{\frac{\sum_i f_{i1}\overline{P_{i1}A_1}^2 + \sum_i f_{i2}\overline{P_{i2}A_2}^2 + \cdots\cdots + \sum_i f_{in}\overline{P_{in}A_n}^2}{f_1+f_2+\cdots\cdots+f_n}}$$

$$=\sqrt{\frac{\sum_i f_{i1}(X_i-A_1)^2+\sum_i f_{i2}(X_i-A_2)^2+\cdots\cdots+\sum_i f_{in}(X_i-A_n)^2}{f_1+f_2+\cdots\cdots+f_n}}$$

$$=\sqrt{\frac{1}{N}\sum_{i,k} f_{i,k}(X_i-A_k)^2}\ 。$$

此量稱爲由直線 l 所起之橫的標準差，得用以測量平面上各點與此直線在橫的方向上之撒布情形。

依上述知在一直線上欲使其標準差變爲最小時，只須將其測量偏差之基準放在算術平均數上卽得。現依同樣理由得考察平面上所有各點對於橫的方向上之撒布情形，依此特採用將橫的標準差成爲最小之直線爲基準，以議論一切。此爲基準之直線，稱爲橫的最小偏差線。此橫的最小偏差線之作成方法，旣與橫的迴歸線之由連接各橫行上之最小偏差點以作成之方法相類似，依此知其與上述二項條件相符合，故得將此橫的最小偏差線定爲橫的迴歸直線。

同樣，一直線 l 與 $X_1, X_2, \cdots\cdots X_m$ 所當縱線之交點爲 $B_1, B_2, \cdots\cdots B_m$ 時，數量

$$\sqrt{\frac{1}{N}\sum_{i,k} f_{i,k}(Y_k - B_l)^2},$$

得稱爲由直線 l′ 所起之縱的標準差，將此量成爲最小時之直線 l′，稱爲縱的最小偏差線，且此縱的最小偏差線得定爲縱的迴歸直線。

此後對該二直線，不再使用最小偏差線(line of least deviation)名稱，直以橫的及縱的迴歸直線以名之。

第五節　迴歸直線之方程式

實際上依 X，Y 軸以求迴歸直線之方程式時，其計算非常複雜。現爲便利起見，特將座標原點移至相關中心 M (M_x, M_y) 上，並依

$$X - M_X = x$$

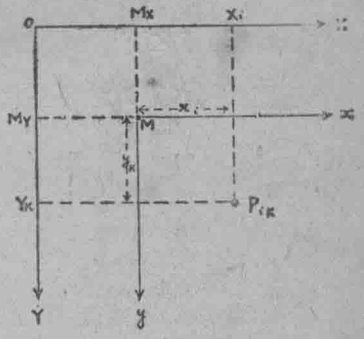
$$Y - M_Y = y$$

以採用新座標(x,y)。其任意一點 $P_{i,k}$ 之新座標成爲

$$x_i = X_i - M_X,$$

$$y_k = Y_k - M_Y.$$

其座標間之關係情形如第 130 圖，其由相關中心 M 至各點之橫偏差總和與縱偏差總和，依相關中心之性質，知皆等於零，即



第 131 圖

$$\sum_{i,k} f_{i,k}\, x_i = \sum_i f_i' x_i = 0,$$

$$\sum_{i,k} f_{i,k}\, y_k = \sum_k f_k y_k = 0.$$

其形式上之證明如次：

$$\sum_{i,k} f_{i,k}\, x_i = \sum_{i,k} f_{i,k}\,(X_i - M_X)$$

$$= \sum_{i,k} f_{i,k}\, X_i - M_k \sum_{i,k} f_{i,k}$$

$$= \sum_{r,k} f_{i,k}\, X_i - \frac{\sum_{i,k} f_{i,k}\, X_i}{\sum_{i,k} f_{i,k}} \sum_{i,k} f_{i,k}$$

$$= \sum_{i,k} f_{i,k}\, X_i - \sum_{i,k} f_{i,k}\, X_i$$

$$= 0_{\circ}$$

同樣得證明

$$\sum_{i,k} f_{i,k}\, y_k = 0_{\circ}$$

在此等預備之下以求迴歸直線之方程式：

設任意一直線 l 之方程式爲

$$x = ay + b,$$

此處是表示 Y 之變化上所起對應 X 之平均的變化情形，故取如上式。

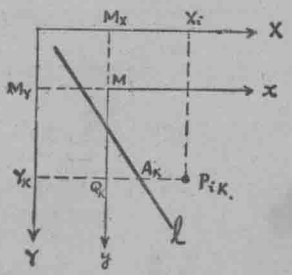
橫軸 Y_k 與 l 相交點爲 A_k，與 y 軸相交點爲 Q_k 時，

$$\overline{MQ_k} = y_k,$$

但 A_k 在直線 l 上，故得

$$\overline{Q_kA_k} = ay_k + b,$$

由此得



第 132 圖