數學傳播　31卷1期, pp. 49-67

統計裡的關係

黃文璋

一. 前言

先看底下一則“給你笑笑”, 作者是方衍濱, 原文刊載於民國94年11月28日中國時報浮世繪 (E6版)：

最近參加一個優良出版品頒獎典禮, 在活潑熱鬧的踢踏舞後,
男主持人說：「今天的舞蹈與出版品有很高的相關性?」
女主持人不解：「有什麼關係?」
男主持人說：「都是腳踏實地。」
女主持人說：「再掰吧!」
男主持人接著說：「都是一步一腳印。」
女主持人說：「還有呢?」
男主持人想了一下說：「最重要的是都跟得上時代的腳步!」
話一落下, 全場掌聲響起。

民國94年12月18日, 中國時報有一則關於眼科醫師梁中玲, 發表“高度近視家族史對子女近視年齡與度數的影響”論文之報導。新聞中指出, 高度近視的父母, 所生兒女出現高度近視的機率, 為一般人之7.7倍, 且平均11歲開始出現近視。還說這是國內首度研究證實視力和基因遺傳的直接相關性。

一般多認為高度近視, 可能與遺傳基因或環境有關, 梁中玲遂蒐集887名17歲至45歲的高度近視者, 進行上述研究。

我們常提到關係, 或說相關、關連。例如, 有沒有關係? 這兩件事是相關的, 統一教與基督教存在高度的關連性等。有一首陳慧琳主唱的歌, 歌名就叫完美關係。比較正式的關係有：

親子關係、手足關係、師生關係、婚姻關係、婆媳關係、家庭關係、
兩性關係、人際關係、勞資關係、自我關係、公共關係、國際關係等。

我國與美國間有“台灣關係法”, 對大陸有“台灣地區與大陸地區人民關係條例”, 在日本設有“亞東關係協會”。國民中小學九年一貫課程綱要中, 有“分段能力指標與十大基本能力之關係”。不論是人、事、物、群體、概念, 任兩者之間, 都可有某種關係。警方及調查人員辦案時, 也常需由蛛絲馬跡中, 找出其中兩兩間的各種關係。大家在物理課程中也學過不少關係。例如, 自由落體, 時間與速度間之關係。或如愛因斯坦著名的 $E=MC^2$ 的公式。Google 的搜尋演算法, 將其他網站連結到該網頁的數目列入計算, 這對特定查詢來說, 爲一重要的相關性指標。可以這樣講, 在數學及科學中 (包含自然、人文及社會科學), 以及在生活裡, 人們常在探討各種情況下, 二者間的關係。光是數學中的關係, 可說就已不勝枚舉：

> 等價關係、對稱關係、相等關係、全等關係、相似關係、平行關係、垂直關係、相依關係、邊與角的關係、直線和圓的位置關係、根與係數的關係, $\cdots$ 。

任二數、形、集合、命題間, 總要看其間有何影響、牽涉, 或連帶作用。

有兩種關係是數學中常出現的, 其一爲*函數關係*, 其二爲*因果關係*。

在數學中*函數*(function) 的意義大致是這樣的：給二集合 A 與 B, 由 A 至 B 的一對應(即每一 A 中的元素 x, B 中恰有一元素 y 與 x 對應), 便稱爲一由A至 (或稱*映至*)B的函數。A 稱爲此函數之*定義域*, B稱爲*對應域*。有時又可把函數想成一台機器, 放進某一原料x, 經過一些作用, 出來某一產品 y。

可看出函數的定義並未有太多限制。現代數學的發展, 大約始自十七世紀, 自那時起, 數學家在討論數學時, 常將問題設法以函數的形式來描述。在機率與統計裡, 也處處藉用函數的概念。例如, 對一隨機現象 (即事先不能預知結果的試驗、實驗或觀測中的某些量), 常藉*隨機變數*來描述, 而隨機變數其實就是函數：一個由*樣本空間*映至實數的函數。

對一函數關係, 常以

$$y=f(x) \tag{1}$$

表之。其中 f 爲函數的名字, 對定義域中的一個x, 經過 f 的作用後, 得到對應域中的 y。由 (1) 式可看出數學的特色之一：常可以簡潔的符號來替代一複雜或冗長的敍述。

函數關係可說是一種很強的關係。給定 x, 經由 f, 所對應的 y 便完全確定。因果關係則是在某*假設*(或說*前提*) 下, 可導致某*結論*成立。常以*命題*若 p 則 q 來描述。此命題何時爲眞? 就是p成立時, 的確可導致 q 成立。當此命題爲眞時, p 稱爲*因*, q 稱爲*果*。有時 q 成立, 也會導致 p 成立。例如, 若爲全等三角形 (p), 則三對應邊等長 (q)。反之, 若三對應邊等長 (q), 則爲全等三角形 (p)。此時 p 與 q *互爲*因果關係。另外, 若爲全等三角形 (p), 則三對應角相等 (q), 但三對應角相等 (q), 卻不一定爲全等三角形 (p)。

有些函數關係也是一種因果關係。令 $y = x^2$, 這是指定出來的函數, 並無法說明白其中有何因果關係。但圓面積爲半徑 r 的函數, 即 $A(r) = \pi r^2$, 其中 $A(r)$ 表半徑 r 的圓面積, π 爲圓周率。此函數關係, 乃是經過證明而得到的。即有命題

$$\text{若圓的半徑爲 } r, \text{ 則其面積} = \pi r^2.$$

當然我們知道對本例亦有

$$\text{若圓的面積爲 } A, \text{ 則其半徑} = \sqrt{\frac{A}{\pi}}.$$

又並非每一因果關係皆爲一函數關係。例如, 若 $x > 0$, 則 $x + 3 > 0$, 便不是一函數關係。

口語裡的關係, 有時沒有太嚴格的定義, 如本節一開始那則給你笑笑中的 "舞蹈與出版品" 的關係。數學裡的各種關係, 定義當然是很明確的。在科學裡, 常在討論隨機現象。對二隨機變數, 我們常想建立其間之關係。但二隨機變數間, 往往沒有像數學中那麼明確的關係。在數學裡若覺得某種關係是對的, 就要去證明。數學中自有一套邏輯推演, 有被認爲究竟怎樣才算得證的程序。但在科學裡, 往往無法如數學中, 天衣無縫的證明一關係成立。前面所引有關近視的報導, 其中的證實, 可不是如數學中, 一步步推導證出視力與基因遺傳有某種關係。那這裡證實二字到底是什麼意思呢? 科學中的證實, 常是經由統計裡的假設檢定(hypothesis testing) 所得之推論, 可參考黃文璋 (2005) 一文。

本文便是要探討隨機現象中, 諸如獨立、相關、關連、無關等關係的題材。

二. 獨立

統計裡常在做預測或估計的工作。收集資料後, 對未來做預測, 或估計某些未知的量。有些情況的預測, 統計學家往往束手無策。例如, 下一期樂透彩頭獎號碼, 雖知道過去每一期之頭獎號碼, 正規的統計學家, 大約不會去做此預測, 因爲各期所開出之頭獎號碼爲相互獨立(mutually independent, 簡稱獨立)。媒體上統計出氣較旺的號碼(常出現), 及氣較弱的號碼(久未出現), 你該簽何者呢? 好像各有道理。事實上, 因每組號碼出現的機率每次均相同, 知道過去的資料, 對預測未來, 並沒有幫助。所以統計學家對這種預測絲毫不顯得高明。另外, 在諸如估計銅板出現正面的機率 p, 通常的作法是持續地投擲 (假設n次), 統計出現幾個正面 (假設 k 次), 然後以 k/n 來估計 p。這其中一個常做的假設是, 各次投擲所出現的結果是"獨立且有共同分佈"(independent and identically distributed, 簡稱 iid)。在很多實際的情況, 樣本常可視爲或設爲獨立。不少統計的理論, 其基本假設往往是樣本爲 iid。

日常生活裡常聽到"獨立"二字 : 主權獨立、經濟獨立, 甚至金雞獨立。統計裡的獨立, 又稱隨機獨立, 與前述獨立的意義不盡相同。簡單地講, 對二事件 A, B, 若給定 B 發生, A 發生

之機率不受影響, 則稱A與B獨立。即若

$$P(A|B)=P(A) \tag{2}$$

則稱 A 與 B 獨立。通常寫成 $P(A|B)$ 時, 要求 $P(B)>0$, 事件 B 總要有可能發生 (即 $P(B)>0$), 給定 B 才有意義。由於依定義

$$P(A|B)=\frac{P(A\cap B)}{P(B)},$$

由 (2) 式引伸出, 若

$$P(A\cap B)=P(A)P(B), \tag{3}$$

則稱 A 與 B 獨立。我們便以 (3) 式當做 A 與 B 獨立的定義, 在此式裡, 就不要求 $P(B)>0$。當然如果 $P(B)>0$, 此時 (2) 式與 (3) 式便無差別。

交集表同時發生。(3) 式顯示, 二事件 A 與 B 獨立, 若且唯若 A,B 同時發生的機率, 等於A,B各自發生的機率之乘積。假設骰子爲公正, 即每面出現之機率皆爲$1/6$。投擲第一個骰子, 會得到點數1(事件 A) 之機率爲 $1/6$, 投擲第二個骰子, 會得到點數2(事件 B) 之機率亦爲 $1/6$。那各投擲一次, 第一個得到1, 第二個得到2之機率爲何? 如果兩次投擲爲獨立, 即互不受影響, 顯然答案爲 $1/6\cdot 1/6=1/36$。即有 $P(A\cap B)=P(A)P(B)$。

有時會有 n 個事件, n 個事件獨立的定義爲何?

設有事件 $A_1,A_2,\ldots,A_n, n\geq 3$。若對 $\forall\ 1\leq k\leq n$, 及 $1\leq i_1<i_2<\cdots<i_k\leq n$,

$$P(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k})=P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k}), \tag{4}$$

則稱 $A_1,A_2,\ldots,A_n$ 爲獨立事件, 或說 $A_1,A_2,\ldots,A_n$ 相互獨立 (獨立)。

可看出要驗證 $A_1,A_2,\ldots,A_n$獨立, 就要驗其中任意挑出之 k 個事件之交集的機率, 等於機率之乘積, 而不只是驗證

$$P(A_1\cap A_2\cap\cdots\cap A_n)=P(A_1)P(A_2)\cdots P(A_n).$$

有些事件不是那麼容易看出獨立或不獨立, 見下例。

例 1:　設一袋中有9張紙牌, 分別寫著 (a,b,c), (a,c,b), (b,a,c), (b,c,a), (c,a,b), (c,b,a), (a,a,a), (b,b,b) 及 (c,c,c) 等。隨機抽取一張, 假設每一張紙牌被取中之機率均爲 $1/9$。令事件 A_k 表抽中的紙牌上所寫之第 k 個字母爲 a, $k=1,2,3$。如 $A_1=\{(a,b,c), (a,c,b),(a,a,a)\}$。顯然

$$P(A_1)=P(A_2)=P(A_3)=\frac{1}{3},$$
$$P(A_1\cap A_2)=P(A_1\cap A_3)=P(A_2\cap A_3)=\frac{1}{9}.$$

因此 A_1 與 A_2, A_1 與 A_3, A_2 與 A_3 皆獨立。但因

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{9} \neq P(A_1)P(A_2)P(A_3),$$

故 A_1, A_2, A_3 不爲獨立事件。

事實上, 若 A_1 與 A_2 均發生, 則 A_3 必發生, 所以 A_3 與 $A_1 \cap A_2$ 不獨立, 因此 A_1, A_2, A_3 不相互獨立。

二隨機變數獨立, 表知道其中之一的值, 對另一變數毫無影響。正如由 (2) 式演變至 (3) 式, 對二離散型之隨機變數 X, Y, 我們定義其獨立, 若對所有可能取的值 a, b,

$$P(X = a, Y = b) = P(X = a)P(Y = b). \tag{5}$$

至於對一般的隨機變數X, Y, X與Y獨立, 表滿足

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),\ \forall x, y \in R. \tag{6}$$

這是較一般的以分佈函數來定義, 如果機率密度函數(probability density function) 存在, 也可以機率密度函數來表示。設 $f(x, y)$ 表 X, Y 之聯合機率密度函數, $f_X(x)$, $f_Y(y)$ 分別表 X, Y 之邊際機率密度函數, 則 X 與 Y 獨立, 若且唯若

$$f(x, y) = f_X(x)f_Y(y),\ \forall x, y \in R. \tag{7}$$

對於 n 個隨機變數 $X_1, \ldots, X_n$, 其獨立的定義爲滿足

$$\begin{aligned} &P(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= P(X_1 \leq x_1) \cdots P(X_n \leq x_n),\ \forall x_1, \ldots, x_n \in R, \end{aligned} \tag{8}$$

或

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n),\ \forall x_1, \ldots, x_n \in R. \tag{9}$$

其中 f 爲 $X_1, \ldots, X_n$ 之聯合機率密度函數, f_i 爲 X_i 之邊際機率密度函數, $i = 1, \ldots, n$。

本節一開始提到, 在做估計時, 常假設樣本爲 iid。例如, 欲估計某銅板出現正面之機率 p, $0 < p < 1$。依序投擲 n 次, 得到 $X_1, \ldots, X_n$, 其中 $X_i = 1$ 或 0, 就依第 i 次得到正面或反面。很自然地, 以樣本平均(sample mean)

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n},\ n \geq 1,$$

來估計p。弱大數法則(weak law of large numbers) 告訴我們, $n \to \infty$時, $\overline{X}_n$ 會機率收斂(converges in probability) 至p。即

$$\lim_{n\to\infty} P(|\overline{X}_n - p| < \varepsilon) = 1,\ \forall\varepsilon > 0. \tag{10}$$

簡單地講, n 很大時, $\overline{X}_n$ 接近 p 之機率很大。這其中 $X_1, \ldots, X_n$ 爲 iid 是一關鍵。如果投擲不一樣的銅板, 每次出現正面的機率不同, 則 (10) 式通常就不成立了。如果各次投擲不獨立, (10) 式也就不見得成立。例如, 有人並未眞的投擲n次, 而是看到第一次出現的結果 X_1 就照抄, 因而得到 $X_2 = \cdots = X_n = X_1$。如此 $\overline{X}_n = X_1,\ \forall n \geq 1$。而 X_1, 不是 1 就是 0, 故不論 n 多大, $\overline{X}_n$ 當然不會接近 p。

三. 迴歸

獨立是一種很特殊的關係, 就是毫無關係。即你怎麼樣, 對我沒有影響。有些事件或隨機變數並不獨立, 不獨立才易做預測。準備考試要先做考古題, 因知道下次的考題, 與過去的考題多少有些關連。送女朋友禮物, 要先了解她的愛好。但如果她喜怒無常, 打聽的結果可能幫助不大。第一次考試成績與第二次考試成績, 兒子身高與父親身高, 明天氣溫與今日氣溫, 這些情況中涉及的兩個量, 很可能都不獨立。既然不獨立, 其間便多少有關係。如何表示其關係呢?

對隨機變數討論關係, 乃指機率上的關係。獨立就是機率分佈彼此沒有任何關連, 不獨立就是機率分佈有些關連。當 X, Y 二隨機變數不獨立, 一個可說是最簡單的情況, 就是 X 與 Y 有函數關係。如

$$Y = g(X), \tag{11}$$

即知道 X 之值後, Y 就完全決定了。

對二隨機變數 X, Y, 能夠知道其聯合分佈函數

$$F(x,y) = P(X \leq x, Y \leq y),\ \forall x, y \in R,$$

或者聯合機率密度函數 $f(x,y)$, 則 X 與 Y 間如何互動, 或者說二者間之關係就完全知道了, 這可說是非常淸楚的關係。例如, 由 $f(x,y)$ 我們可求出條件機率密度函數

$$f(y|x) = \frac{f(x,y)}{f_X(x)}, \tag{12}$$

其中如果是連續型變數, 則

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy. \tag{13}$$

若 X 與 Y 有如 (11) 式之函數關係, 則給定 $X=x, Y=g(x)$ 之機率爲 1。即

$$P(Y=g(x)|X=x)=1,$$

此時

$$f(y|x)=\begin{cases}1, \text{若}\, y=g(x),\\ 0, \text{若}\, y\neq g(x).\end{cases}$$

$f(y|x)$ 表給定 $X=x$ 之下, Y 之機率密度函數。只要 X 與 Y 不獨立, $f(y|x)$ 便不等於 $f_Y(y)$。即 Y 之條件分佈, 隨著給定 X 不同的値而不同。由於期望值爲關於隨機變數之一重要的量, 因此諸如條件期望值 (仍設 X, Y 爲連續型變數)

$$E(Y|X=x)=\frac{\int_{-\infty}^{\infty} yf(x,y)dy}{\int_{-\infty}^{\infty} f(x,y)dy}, \tag{14}$$

便是對於二變數 X, Y 常會討論的題材。曲線 $y(x)=E(Y|X=x)$, 便稱做*迴歸曲線*(regression curve), 代表 Y 對 X 之*迴歸*。當然亦可定義 X 對 Y 之迴歸曲線 $x(y)=E(X|Y=y)$。當迴歸曲線爲直線, 此時便稱*線性迴歸*(linear regression)。

隨機現象裡, 通常少有如 (11) 式, Y 就是 X的函數。但若有 (11) 式的關係, 對預測而言, 當然是最完美的。像是知道父親身高 (X), 兒子身高 (Y) 也就確定。退而求其次, 如果能有如下關係：

$$Y=g(X)+\varepsilon, \tag{15}$$

其中 ε 代表誤差項, 大約也會令人滿意。通常會對 ε 做一些假設。如因是誤差, ε 有時爲正, 有時爲負, 平均來說, 似該爲0, 所以假設 $E(\varepsilon)=0$。有時爲了簡便, 或是採信高斯的*誤差理論*, 而假設 ε 有常態分佈。g 如果是一很簡單的函數, 如線性函數 $g(x)=ax+b$, 就會更令人滿意。依所收集到的數據 $(x_1,y_1),\ldots,(x_n,y_n)$, 以估計 a, b。這方面的討論, 就是統計裡的*迴歸分析*(regression analysis), 是很基本且重要的題材。

當 X, Y 之聯合分佈知道, 有時想找一適當的函數g, 並以 $g(X)$ 來預測 Y。也就是 X 爲我們所知的資訊, 而Y便是要預測的值。對 Y 之預測值 $g(X)$, 我們自然希望要愈接近 Y 愈好。但怎樣是接近? 一個代表誤差的量是 $g(X)$ 與 Y 之差的絕對值 $|g(X)-Y|$, 即*絕對誤差*。由於此爲一隨機變數, 期望值 $E(|g(X)-Y|)$ 爲代表其大小的一個量。但涉及絕對值的積分或是和, 通常較難計算。只要想想積分 $\int_{-\infty}^{\infty}|x^2+3x-e^x|dx$ 就知道了。計算前要先去掉絕對值, 而 x^2+3x-e^x 何時會正或負, 並不好決定。所以通常考慮誤差平方 $(g(X)-Y)^2$ 之期望值 $E((g(X)-Y)^2)$, 或說*均方差*(mean squared error, 簡稱 MSE)。這就像對一隨機

變數U, 以 $E((U-E(U))^2)$ 表其變異數, 並稱其正的平方根 $[E((U-E(U))^2)]^{1/2}$ 為標準差, 用來量測 U 與其期望值 $E(U)$ 偏離的大小。若有一 Y 之預測值 $g(X)$, 使得其 MSE 最小, 我們便稱此為 Y 之最佳MSE 預測值。MSE是統計學中, 常用來表示誤差的一個量。

可以證明條件期望值 $g(X)=E(Y|X)$, 為唯一的 Y 之最佳 MSE 預測值。但若只想簡單地以一常數來預測 Y, 則 $g(X)=E(Y)$ 可使預測之 MSE 最小。這也解釋對一隨機變數, 何以期望值為其一常用的代表值。因在均方差最小的意義下, 此為隨機變數之最佳常數預測值。若想以一線性函數 $g(X)=aX+b$ 來預測 Y, 則當 $E(X)$, $E(Y)$, $\mathrm{Var}(X)$ 皆存在, 可得

$$a=\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \quad b=E(Y)-aE(X), \tag{16}$$

其中

$$\begin{aligned}\mathrm{Cov}(X,Y)&=E((X-E(X))(Y-E(Y)))\\&=E(XY)-E(X)E(Y),\end{aligned} \tag{17}$$

表 X 與 Y 之共變異數(covariance, 又稱協方差)。上述這些最佳預測值如何求出, 可參考黃文璋 (2003a)3.5 節。

四. 相關係數

知道 X,Y 的聯合分佈, 對 X,Y 如何互動雖然完全掌握, 但有時我們只想粗略地了解 X 與 Y 間的關係有多密切, 即想以一簡單的量來描述其關係。此正如對一場考試, 我們不見得對全部考生的成績感興趣, 平均值、標準差反而較想知道。因為此二值, 能讓人快速地了解全部考生成績之集中及散佈情況。相關係數(correlation coefficient, 或只稱 correlation) 的概念就因此產生。

如果二隨機變數不獨立, 則二變數間便有關係, 此關係可能強可能弱。如果是獨立, 則關係當然是最弱的。以水為例, 令 X 表水的體積, Y 表水的重量。顯然 X 與 Y 之關係很強。如果取樣多次, 如n次, 將所得的數據 $(x_1,y_1),\ldots,(x_n,y_n)$ 畫在 x-y 座標平面上, 則所有的點很可能落在一直線上或直線的附近。這是因為純淨的水重量與體積有線性關係。有些數據不在直線上, 可能是因量測的誤差, 或水質不純所致。其次, 若以 X 表某人之身高, Y 表其體重, X 與 Y 顯然亦有關係, 但可能不像水的重量與體積間之關係那麼強。量測 n 個不同的人所得之 $(x_1,y_1),\ldots,(x_n,y_n)$, 大約不會形成一直線, 雖然我們仍會預期在 x-y 座標平面上, 圖形是向右上增長, 即較大的 x 有較大的 y 之傾向。共變異數及相關係數, 都可用來度量兩隨機變數關係(特別是線性關係) 的強弱。

在本節中, 爲了簡便, 令 $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$, 如果這些量存在的話。又所有結果的證明, 皆可參考黃文璋 (2003b)3.5節。首先當 $0 < \sigma_X^2, \sigma_Y^2 < \infty$, 定義 X, Y 之相關係數爲

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{18}$$

$\mathrm{Cov}(X,Y)$ 可以 σ_{XY} 表之, $\rho(X,Y)$ 有時以 ρ_{XY}, 或簡單地以 ρ 表之。

當 σ_X 或 $\sigma_Y = 0$, 此時 $\mathrm{Cov}(X,Y) = 0$, 至於 $\rho(X,Y)$ 則不定義。由相關係數之定義知, 只要我們提到相關係數, 皆隱含二隨機變數之期望值及變異數皆存在, 且變異數皆不爲0。

顧名思義, 共變異數量測兩個隨機變數同時變化的情況。正如變異數乃量測一隨機變數變化的大小。自己跟自己的共變異數就是變異數, 即 $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$。如果較大的 X 傾向於伴隨較大的 Y, 且較小的X傾向於伴隨較小的 Y, 則 $\mathrm{Cov}(X,Y)$ 將是正的。因此若 $X > \mu_X$ 時, 較可能有 $Y > \mu_Y$, 則 $(X-\mu_X)(Y-\mu_Y)$ 較可能爲正; 且若 $X < \mu_X$ 時, 亦較可能有 $Y < \mu_Y$, 則 $(X-\mu_X)(Y-\mu_Y)$ 也將較可能爲正。如此一來, $\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) > 0$。但若較大的$X$傾向於伴隨較小的$Y$, 且較小的 X 傾向於伴隨較大的Y, 則 $(X-\mu_X)(Y-\mu_Y)$ 便較可能爲負, 如此一來, $\mathrm{Cov}(X,Y) < 0$。故 $\mathrm{Cov}(X,Y)$ 之正負, 反映 X, Y 增長方向之相同或相反。若以X, Y分別表父親的身高及兒子的身高, 我們會預期 $\mathrm{Cov}(X,Y)$ 爲正。若以 X, Y 分別表成人男子的體重及跳高的高度, 我們會預期 $\mathrm{Cov}(X,Y)$ 爲負。

可以證明當 X 與 Y 獨立時, $\mathrm{Cov}(X,Y) = 0$。不過 $\mathrm{Cov}(X,Y) = 0$ 時, X 與 Y 卻不一定獨立。另外, $\rho(X,Y)$ 與 $\mathrm{Cov}(X,Y)$ 之正負符號相同, 且當 σ_X及σ_Y 皆存在時, $\rho(X,Y) = 0$, 若且唯若 $\mathrm{Cov}(X,Y) = 0$。當 $\rho(X,Y) = 0$, 則稱 X 與 Y無相關(uncorrelated)。

做爲量測二隨機變數 X 及 Y 之變化情況, 共變異數有一缺點, 就是其值與量測 X, Y 之尺度有關。例如, X 以公升爲單位, Y 以公斤爲單位, 跟 X 以公撮 (立方公分) 爲單位, Y 以公克爲單位, 則前者求出之共變數爲後者之 10^{-6} 倍。相關係數 ρ 便可消除這種困擾, 它永遠取值在區間 $[-1,1]$。$|\rho|$ 接近1時, 顯示 X 與 Y 有較強的線性關係, $|\rho|$ 接近0時, 顯示 X 與 Y 的線性關係較弱, 而 $|\rho| = 1$ 時, 表 X 與 Y 有完美的線性關係。

可以證明, 對二隨機變數 X, Y,

$$\rho(X,Y) = \mathrm{Cov}(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}) = E(\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}). \tag{19}$$

我們知道, 對一隨機變數 X, 所謂標準化就是將 X 減去期望值, 再除以標準差, 而得 $(X-\mu_X)/\sigma_X$。由 (19) 式知, 相關係數即二標準化後的隨機變數之共變異數。又對任二隨機變數

X, Y, 及常數 a, b, c, d,

$$\operatorname{Cov}(aX+b, cY+d) = ac\operatorname{Cov}(X,Y), \tag{20}$$
$$\rho(aX+b, cY+d) = \frac{ac}{|ac|}\rho(X,Y), ac \neq 0. \tag{21}$$

(20) 式印證我們前面所說的共變異數與所取的尺度 (指 a, c 的效應) 有關。不過與座標的平移 (指 b, d 的效應) 無關。(21) 式指出, 只要尺度之改變, 並未使 X, Y 之值的方向變成相反 (即 $ac > 0$), 則相關係數不變。

例 2: 設 X, Y 之聯合機率密度函數爲

$$f(x,y) = 1, 0 < x < 1, x < y < x+1.$$

則 X, Y 之邊際機率密度函數分別爲

$$f_X(x) = 1, 0 < x < 1,$$
$$f_Y(y) = \begin{cases} y, & 0 < y < 1, \\ 2-y, & 1 \leq y < 2. \end{cases}$$

則 $\mu_X = 1/2$, $\sigma_X^2 = 1/12$, $\mu_Y = 1$, $\sigma_Y^2 = 1/6$。又 $E(XY) = 7/12$。因此

$$\operatorname{Cov}(X,Y) = E(XY) - \mu_X\mu_Y = \frac{7}{12} - \frac{1}{2}\cdot 1 = \frac{1}{12},$$
$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{1/12}{\sqrt{1/12}\sqrt{1/6}} = \frac{\sqrt{2}}{2} \doteq 0.707.$$

例 3: 設 X 與 Z 爲二獨立的隨機變數, 且 X 有 $\mathcal{U}(0,1)$ 分佈, Z 有 $\mathcal{U}(0,1/10)$ 分佈。令 $Y = X + Z$。利用變數代換, 可得 X, Y 之聯合機率密度函數爲

$$f(x,y) = 10, 0 < x < 1, x < y < x + \frac{1}{10},$$

如此便可求出 $\operatorname{Cov}(X,Y)$ 及 $\rho(X,Y)$。不過因 Y 爲 X 與 Z 之和, 且 X 與 Z 獨立, 我們可經由 X 與 Z, 如下求出 $\operatorname{Cov}(X,Y)$ 與 $\rho(X,Y)$。

$$E(X) = \frac{1}{2}, \quad \operatorname{Var}(X) = \frac{1}{12},$$
$$E(Y) = E(X+Z) = E(X) + E(Z) = \frac{1}{2} + \frac{1}{20} = \frac{11}{20},$$
$$\operatorname{Var}(Y) = \operatorname{Var}(X+Z) = \operatorname{Var}(X) + \operatorname{Var}(Z) = \frac{1}{12} + \frac{1}{1{,}200} = \frac{101}{1{,}200}.$$

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y) \\
&= E(X(X+Z)) - E(X)(E(X)+E(Z)) \\
&= E(X^2) + E(XZ) - (E(X))^2 - E(X)E(Z) \\
&= E(X^2) - (E(X))^2 \\
&= \mathrm{Var}(X) = \frac{1}{12}.
\end{aligned}$$

此處用到因 X 與 Z 獨立, 所以 $E(XZ) = E(X)E(Z)$。由此又得

$$\rho(X,Y) = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}}\sqrt{\frac{101}{1{,}200}}} = \frac{10}{\sqrt{101}} \doteq 0.9950.$$

與例2比較, 在例3中, X 與 Y 的相關係數很接近1。我們將兩例中使 $f(x,y)$ 爲正之 (x,y) 的圖形繪出, 見圖1。

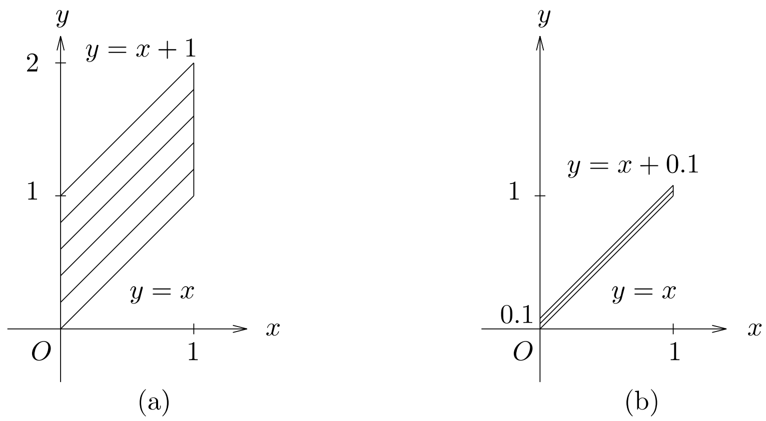


圖1. (a) 例2中使$f(x,y) > 0$之區域, (b) 例3中使$f(x,y) > 0$之區域。

圖1(a) 及 1(b), 顯示 X 與 Y 大致有線性的關係, 但在圖 1(b) 中, 線性關係較強 (區域較窄)。也可以另一方式來看此線性關係的強弱。在例3中, 因 $Y = X + Z$, 故給定 $X = x$, Y 有 $\mathcal{U}(x, x+1/10)$ 分佈。而在例2中, 也可驗證給定 $X = x$, Y 有 $\mathcal{U}(x, x+1)$ 分佈。故知道 $X = x$ 後, 在例3中, 比在例2中, 提供較多關於Y之資訊 (前者誤差不超過0.1, 後者誤差不超過 1)。因此在例3中, X 與 Y的相關係數較大。

相關係數可用來量測二隨機變數之關連程度(degree of association)。但它作爲一個指標, 主要是反映二隨機變數之分佈的線性關係之強度及正負符號。因此相關係數爲 0, 僅表示二隨機變數之線性關係很低, 而非表示二變數機率上無關(probabilistically unrelated), 也就是二

變數不一定獨立。甚至二隨機變數間, 可能有強烈的關係, 但該關係並非線性, 所以相關係數仍可能爲0。見下二例。

例 4: 設 X 與 Z 爲二獨立的隨機變數, 且 X 有 $\mathcal{U}(-1,1)$ 分佈, Z 有 $\mathcal{U}(0,1/10)$ 分佈。令 $Y=X^2+Z$。可看出給定 $X=x$, Y 有 $\mathcal{U}(x^2,x^2+1/10)$ 分佈, 因此 X 與 Y 有很強的關係, 只是並非線性。又可求出 X,Y 之聯合機率密度函數爲

$$f(x,y)=5,\ -1<x<1, x^2<y<x^2+\frac{1}{10}.$$

如此便可求出 $\mathrm{Cov}(X,Y)$ 及 $\rho(X,Y)$。不過如同例3, 我們以下述方式求之。由於 X 有 $\mathcal{U}(-1,1)$ 分佈, 故

$$E(X)=E(X^3)=0.$$

又因 X 與 Z 獨立, 故

$$E(XZ)=E(X)E(Z)=0.$$

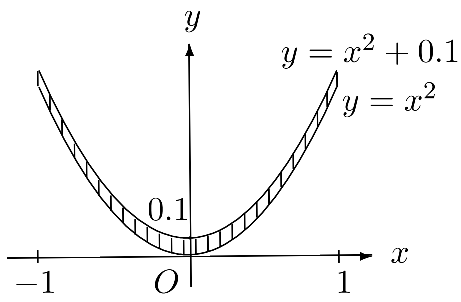


圖2. 例4中使 $f(x,y)>0$ 之區域。

因此

$$\begin{aligned}\mathrm{Cov}(X,Y)&=E(X(X^2+Z))-E(X)(E(X^2+Z))\\&=E(X^3)+E(XZ)=0,\end{aligned}$$

且 $\rho(X,Y)=0$。

例 5: 設X有$\mathcal{U}(-1,1)$分佈。令$Y=X^2$, $Z=2X+3$, $W=-2X+3$。則

$$E(X)=0,\quad E(XY)=E(X^3)=0,$$

故

$$\mathrm{Cov}(X,Y)=E(XY)-E(X)E(Y)=0,$$

且$\rho(X,Y)=0$。相關係數是量測線性關係的強弱, 而非量測其他非線性關係之強弱。故此處雖 Y 爲 X 的平方, 二者有一簡單且密切的關係, 仍得到 X 與 Y 無相關。換句話說, 無相關並非不相關。

另外,

$$\begin{aligned}&E(XZ)=E(X(2X+3))=2E(X^2)+3E(X)=\frac{2}{3},\\&E(X)E(Z)=0,\\&\mathrm{Var}(X)=E(X^2)=\frac{1}{3},\quad \mathrm{Var}(Z)=4\mathrm{Var}(X)=\frac{4}{3},\end{aligned}$$

故

$$\mathrm{Cov}(X,Z)=E(XZ)-E(X)E(Z)=\frac{2}{3},$$

且

$$\rho(X,Z)=\frac{\mathrm{Cov}(X,Z)}{\sigma_X\sigma_Z}=\frac{2/3}{\sqrt{1/3}\cdot\sqrt{4/3}}=1.$$

同理可得 $\rho(X,W)=-1$。

X 與 Z 有線性關係, 且增長方向相同, 故相關係數爲1。X 與 W 亦有線性關係, 只是增長方向相反, 所以相關係數爲-1, 一般而言這是對的。

共變異數對了解隨機變數和之變異有很大的幫助。底下先給一二變數和之變異數的公式, 其中當然要假設 $\mathrm{Var}(X)$ 及 $\mathrm{Var}(Y)$ 皆存在。

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y). \tag{22}$$

故若 X 與 Y 爲正相關, 即 $\rho(X,Y)>0$(因此 $\mathrm{Cov}(X,Y)>0$), 則 $X+Y$ 之變異數, 比X,Y變異數之和還大。反之, 若 X 與 Y 爲負相關, 則 $X+Y$ 之變異數, 比 X,Y 變異數之和還小。當負相關時, X 與 Y 一個值較大, 另一值便傾向於較小, 而其和就不至於那麼極端, 因此 $X+Y$ 之變異減小。同理可對 $\mathrm{Var}(X-Y)$ 做討論。

利用著名的史瓦茲不等式 (Schwarz's inequality), 又稱柯西-史瓦茲不等式 (Cauchy-Schwarz's inequality), 可以證明對任二隨機變數 X,Y, 只要 $0<\sigma_X^2,\ \sigma_Y^2<\infty$, 則其相關係數介於正負1之間。即$|\rho(X,Y)|\le 1$, 且 $|\rho(X,Y)|=1$, 若且唯若 $P(Y=aX+b)=1$, 其中 a,b 爲二常數。又 $\rho(X,Y)=1$ 時, $a>0$, $\rho(X,Y)=-1$ 時, $a<0$。由於相關係數的值介於正負1之間, 且當其值等於 $+1$ 或 -1 時, 二變數有一線性關係之機率爲 1, 這便說明了如前所述, 相關係數是用來量測二變數之線性相依情形。當 $|\rho(X,Y)|=1$ 時, 我們稱 X 與 Y 爲完全相關(completely correlated)。

當 $0<\sigma_X^2<\infty$, 我們注意到 $\rho(X,X)=1$。自己與自己的相關係數爲1, 自然是合理的。事實上, 對任二常數a,b, 且 $a\neq 0$, $\rho(X,aX+b)=1$。又讀者可參考黃文璋 (2003b, pp.221-223) 所提供的說明, 以了解爲何採用相關係數來量測二隨機變數的共線性。

如同對一組隨機樣本 $X_1,\ldots,X_n$, $n\ge 2$, 以

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,$$

表樣本平均及樣本變異數 (有時以 $\overline{X}_n, S_n^2$ 表之, 此處省略足標 n)。對兩組隨機樣本 $X_1,\ldots,X_n$, 及 $Y_1,\ldots,Y_n$, 以

$$S_{XY}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}),\tag{23}$$

及

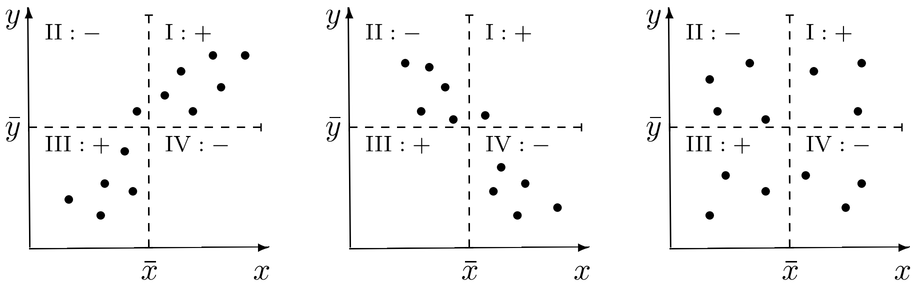
$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2)^{1/2}},\tag{24}$$

分別表樣本共變異數(sample covariance) 及樣本相關係數(sample correlation coefficient)。可看出

$$S_{XY}^2=\frac{1}{n-1}(\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i),\quad r=S_{XY}^2/(S_X^2S_Y^2)^{1/2},$$

其中 S_X^2, S_Y^2 分別為 $X_1,\ldots,X_n$ 及 $Y_1,\ldots,Y_n$ 之樣本變異數。

我們藉圖3來說明。所觀測到的數據 $(x_1,y_1),\ldots,(x_n,y_n)$, 若大部分都落在區域 I 及 III, 則 r 將為正 (正相關); 若大部分都落在區域 II 及 IV, 則 r 將為負 (負相關); 若散佈在 I, II, III, IV等四個區域, 則 r 將接近0(無相關)。



(a) $r>0$ (b) $r<0$ (c) $r\doteq 0$

圖3. (x,y)之散佈情況與相關係數之符號。

最後, 附帶一提, 對二事件A,B, 若滿足$0<P(A),P(B)<1$, 則可如下定義其相關係數$\rho(A,B)$:

$$\rho(A,B)=\frac{P(A\cap B)-P(A)P(B)}{\sqrt{P(A)(1-P(A))P(B)(1-P(B))}}.\tag{25}$$

與隨機變數的相關係數有一些類似的性質。即 $|\rho(A,B)|\le 1$, 且 $\rho(A,B)=1$, 若且唯若 $A=B$; $\rho(A,B)=-1$, 若且唯若 $A=B^c$, 其中 B^c 表 B 之餘集。又 $\rho(A,B)=0$, 若且唯若 A,B 獨立。

五. 相關與因果

如果拜訪客戶之次數 (X) 與客戶購買電腦的數目 (Y), 其相關係數爲正, 且不算太小, 如 0.5, 則顯示廠商拜訪客戶愈頻繁, 客戶買的電腦數有愈多的傾向。這樣一來, 廠商自然受到鼓勵, 會勤於拜訪客戶。二隨機變數若不獨立, 便是有關係。有時會說有關連(association), 或就說相關。雖然在上一節中已介紹, 諸如無相關、正相關、負相關, 及完全相關等, 都有特別的定義, 但人們還是常籠統的以相關一詞, 表示二隨機變數之間有某種關連, 也就是不獨立。相關係數主要是量測二變數間的線性關係。在上一節例4及例5看到, 即使相關係數爲0, 二變數仍可能有密切關係。除了相關係數之大小外, 諸如一個變數隨著另一變數之增大, 而有較高的機率增大, 或較高的機率變小, 也都表二變數相關。相關性爲做決策時之一有效的依據。著名的"尿布與啤酒"事件, 大家可能聽過。Wal-Mart 是美國最大的一家百貨連鎖店, 很擅長購物籃分析 (market basket analysis)。他們檢查顧客的購物清單, 發現啤酒與尿布, 同時出現的比例很高。經過調查發現, 很多家庭主婦會叮嚀先生下班時買尿布。需要尿布的家庭, 總是家中有嬰兒, 而會買尿布的先生, 大抵也較顧家。待在家中照顧嬰兒, 總不能去睡覺或做太專注的事, 於是通常就看看電視以打發時間, 這時啤酒就需要了。弄清楚後, Wal-Mart 遂在各賣場, 將表面上看起來風馬牛不相及的尿布與啤酒擺在一起, 結果尿布與啤酒的銷售量雙雙大增。

網際網路的時代, 使資訊的搜尋及傳遞更便捷。著名的網路書店亞馬遜 (Amazon), 從其"顧客書評"中, 可了解買這本書的顧客, 也買那些書。這種閱讀的相關, 對讀者及出版社, 都是一重要資訊。

不論醫學或商業上, 甚至很多其他的領域, 都常有相關性的研究。例如,

過敏性鼻炎及氣喘的相關性,
大學學測成績與大學成績的相關性,
乳房大小與乳癌發生率的相關性。

乳癌列名婦女十大死因之前幾名, 各種"xx與乳癌的相關性"之研究很多, 連乳房大小都可拿來探討。曾有一則新聞, 標題爲"美政府著手調查潛艦與海豚擱淺事件相關性"。海豚擱淺怪罪潛艦? 原來美國佛羅里達州南部, 有一周發生海豚大量擱淺, 美國政府遂進行調查, 看是否與當地海域, 海軍演習中的潛艦聲納系統有關。但要注意的是, 相關性很高, 不表其中必有因果關係。前言裡那則"給你笑笑", 雖掰出舞蹈與出版品有很高的相關性, 二者間其實並不會有太大關連。千萬不可輕率地對統計上看起來似乎相關的兩個因素, 驟下結論說其間有因果關係。因果關係是否存在, 並非統計理論可以證實。就算利用統計裡的假設檢定, 接受買尿布者, 又買啤酒的比例, 高於又買其他飲料 (如果汁、可樂等) 之假設, 也並未證出買尿布者, 對啤酒的需求較高。因果關係需藉助其他科學方法來判定, 無法只靠統計。見底下二實例。

例 6: 1958年, 著名的統計學家 R. A. Fisher 發表一篇有意思的文章, 標題為 Cigarettes, cancer and statistics(香菸、癌症與統計)。隨後於 1958, 1959 年又在頗負盛名的 Nature 雜誌上發表 Lung cancer and cigarettes? 及 Cancer and smoking 二文。原來老菸槍 Fisher, 對那些抽菸會導致肺癌的研究非常不滿, 他認為其中的所謂"證據", 常是有瑕疵的。例如, 有些數據被刻意隱瞞, 研究報告是經過篩選或刪改等。他覺得政府就是想用一切力量, 讓大家接受抽菸的可怕和危險。讀者可參考葉偉文譯 (2001) 一書第十八章"抽菸會致癌嗎?"一文, 關於此事件之來龍去脈。

時至今日, 儘管反菸人士努力地宣導, 仍有許多癮君子, 何況是五十年前。雖然研究顯示, 肺癌與抽菸的相關性很高, 但如同前面所述, 相關性高並不必然導致其間有因果關係。那些嗜菸者, 當然更不會輕易因一些統計分析而戒菸。就像肉食主義者, 不會因"吃肉致癌的機率是抽煙的1-2萬倍"之報導 (台中榮總醫師王輝明提出的), 而改為吃素。由於愈來愈多的研究指出抽菸對健康不好, 而且抽二手菸也很不好, 再加上反菸團體大力遊說, 我國立法院早於民國86年, 便通過"菸害防治法", 大幅限縮吸菸場所。

例 7: 2004年5月, 在美國紐奧良 (New Orleans, 2005年9月遭卡崔娜颶風重創的美國南部大城), 召開的一研討會上, 印度的一個醫療小組提出報告, 他們從美國人過去半個世紀氣泡飲料的平均消耗, 找到和食道癌罹患率提高的關連。報告中指出, 美國人平均一年飲用氣泡飲料的量, 過去50年增加了450%, 而過去25年來, 美國白人男性食道癌罹患率, 也呈現明顯增加的趨勢。

有人質疑這說不定只是巧合, 或如前所述可能是其他原因所造成。為什麼說可能是巧合? 隨著經濟好轉, 或行銷成功, 不少產品銷售量大增。過去50年不只氣泡飲料的每人每年平均的飲用量, 汽車購買, 旅遊次數, 以及很多其他的消費, 可以想像都是大幅增加。但總不能說都與食道癌有關。不過該研究小組提出科學的理論基礎, 即氣泡飲料會讓胃部膨脹, 導致消化液逆流, 而這是食道癌產生的原因之一。他們的研究還發現, 氣泡飲料的消耗, 和食道癌增加的關連, 為一全球性的趨勢。

兩變數看起來有關連, 有時可能是其中某些潛在變數所造成。著名的辛普生詭論(Simpson's paradox), 也是指因另一變數的介入, 而使兩變數的關係反過來, 見黃文璋 (2006) 例6。下例亦為一種常見的情況。

例 8: 曾有某大學女學生抱怨, 該校研究所對女生較不公平, 因全校研究所女生的總錄取率35%, 明顯較男生的44%低。該校研究所很多, 自其中挑出6個較大的研究所來比較, 可得表1(見 Freedman et al. (1991, p.17))。

6個研究所中, 有 A, B, D, F 等4個研究所, 女生錄取率高於男生; 男生錄取率高於女生的, 僅有 C, E 等2研究所。但若6個研究所一起看, 約有44%的男生被錄取, 卻只有 30% 的女生被錄取, 相差高達14%。此矛盾是如何產生的?

表1. 6個研究所男女生錄取率。

所別		申請	錄取	錄取率
A	女	108	89	82%
	男	825	511	62%
B	女	25	17	68%
	男	560	353	63%
C	女	593	201	34%
	男	325	120	37%
D	女	375	131	35%
	男	417	138	33%
E	女	393	94	24%
	男	191	53	28%
F	女	341	24	7%
	男	373	22	6%
合計	女	1835	556	30%
	男	2691	1197	44%

如果仔細檢查表1, A, B 二研究所較易申請, 且有一半以上的男生 (共1385人) 申請此二研究所, 而 C, D, E, F 四個研究所較難申請, 卻有超過 90% 的女生 (1702人) 申請。換句話說, 男生較多申請較容易進的研究所, 女生較多申請較難進的研究所。因此光由表面上全校研究所總錄取率之高低, 來推測其錄取對女生不公, 並不恰當。事實上, 有可能全校每一研究所都是女生錄取率較高, 但全校一起看, 卻是男生錄取率較高。

這類例子不少。例如, 甲乙兩袋中, 皆有紅白兩種籤, 每支籤可能中獎, 也可能不中獎, 兩袋中皆是紅籤中獎率較高。現將兩袋中的籤混爲一袋。則有沒有可能新袋中, 紅籤中獎率變成較低? 答案是肯定的。甚至新袋中紅籤中獎率, 有可能不到白籤的百分之一 (更小也都可能)。此問題不難, 留給讀者自己舉例好了。

潛在變數的可能存在, 使我們在依相關性做決策時要更小心, 見下例。

例 9: 2006年1月4日有一則“咖啡能減少罹患乳癌風險”的報導。婦女聞乳癌而色變, 若

喝咖啡眞能減少罹患率, 當然是一好消息。

加拿大多倫多大學的研究團隊, 對將近1,700名體內有 BRCA1特殊突變基因的婦女, 探討咖啡和乳癌之間的關係。這個族群的婦女, 在70歲以前, 有0.8的機率會得乳癌。研究結果發現喝咖啡將可降低她們乳癌罹患率。每天若喝1到3杯含咖啡因的咖啡, 罹患率可降低約10%; 喝4到5杯, 可降約25%; 若每天平均喝6杯以上, 乳癌罹患率將減少約75%。研究人員說, 關鍵可能是出在植物性荷爾蒙上, 咖啡中也含有多量的植物性荷爾蒙。

看起來對那些體內有 BRCA1乳癌易感基因的婦女爲一重大消息, 似乎該拼命喝咖啡才是。但咖啡喝太多, 是否會引發其他毛病? 這篇報導可沒說。在準備大喝咖啡前, 還是得多打聽一下 (有人指出喝咖啡若加一包糖及一個奶油球, 便有50大卡熱量, 每天6杯就有300大卡。如果沒有適度運動, 長期下來易發胖。而體內囤積過多脂肪, 恰是乳癌的高危險群!)。

雖然統計分析無法證明因果關係。但統計分析後, 所找出之不同變數間的相關性, 可提供進一步探討的方向。探討結果可能證實其間有因果關係, 或澄清純粹只是巧合, 或其他因素造成。沒有進一步探討, 就冒然論斷因果, 甚至大作文章, 有時會出笑話的。我們以下述事件做爲本文之結束。

2004年3月台灣的總統大選, 引發若干爭議。當年5月, 於總統就職前, 中央研究院有位研究人員, 提出他對廢票率的分析。他指出廢票率與支持度呈現高度的關連性。廢票率愈高的投開票所, 開出的選票中, 陳水扁和呂秀蓮那組較佔上風, 廢票率較低的投開票所, 開票結果, 連戰和宋楚瑜那組較居優勢。他因此斷言廢票決定了此次總統大選的結果, 認爲其中必然有做票。他對自己的研究成果深具信心, 不怕綠營來控告, 並主動放棄抗告的權利。

造成幾天的譁然後, 這件事後來不了了之, 多數人都難以接受這樣的推論。總之, 統計上的相關, 與因果關係, 有時是兩回事。很多情況下, 相關性可做爲決策之依據, 如前述尿布與啤酒的關係之例。無論如何, 將啤酒放在尿布附近, 大約不會有不良後果。但對諸如氣泡飲料和食道癌的關係, 或任何可能會引起很大爭議的事件, 就要有更多佐證後, 才適合提出研究報告, 並給出建議。

致謝辭： 作者感謝評審所提的許多寶貴意見, 使本文更趨完善。

參考文獻

1. 黃文璋 (2003a). 機率論。華泰文化事業股份有限公司, 台北。
2. 黃文璋 (2003b). 數理統計。華泰文化事業股份有限公司, 台北。
3. 黃文璋 (2005). 統計顯著性。數學傳播季刊, 29(4)： 29-38。
4. 黃文璋 (2006). 決策的誤差。數學傳播季刊, 30(3)： 66-79。

5. 葉偉文譯 (2001). 統計改變了世界 (David Salsburg 原著：*The Lady Tasting Tea*)。天下遠見出版股份有限公司, 台北。
6. Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1991). *Statistics*, 2nd ed. W. W. Norton Company, New York.
7. Fisher, R. A. (1958). Cigarettes, cancer and statistics. *Centennial Review*, 2, 151-166.

—本文作者任教於國立高雄大學應用數學系—