萬物有常 世事多變

趙民德
中央研究院統計科行研究所

曾經有一支李壽全的歌, 也許現在還在 KTV 中可以點到, 大概是叫做「我的志願」。歌詞的意思說, 小的時候, 他的志願是做總統, 做發明家, 後來日漸長大, 才知道總統只有一個, 眞正的發明家也不太多。志願在聯考時塡一百多個之後, 也就不成其爲志願了。而當年紀更大了, 才知道天下的事情都是正正常常地運作的, 大多數人都是普通人, 既不是總統, 也不是發明家。雖不是混混噩噩地過日子, 但是快樂是自己定義的; 不快樂也是。因此我們可以悲傷, 可以歡笑, 和普通人一樣。

這個世界本來就是這樣的。不管你自己覺得你是如何的了不起或特立異行, 但大家都是在同樣的環境之下培養出來的, 吃同樣的米, 喝同樣的水, 呼吸同樣髒的空氣, 讀同樣的教本, 看同樣的武俠小說、漫畫和港片。因此不論是在你的身上做甚麼測量—— 比方說腰圍吧 —— 那麼量測出來的數字, 如果放在一大堆類似的人的同樣的測量裡, 就會看不出有多少不同。因爲你是一個群體裡面典型的一員。典型的意思就是說: 大家都是普通人, 你當然也不會特別地特出。

這當然是悲哀, 也是無奈, 但也不是那麼的一定就是壞事。不論甚麼, 我是說不論甚麼, 只要不是強迫性的 (例如二中選一的結果), 只要是可以量測出來 (例如可以到小數二位) 的事物, 大概就是這樣。你不論去看一個羅馬時代的軍團, 或是去看一個成吉斯汗的萬人隊, 大概就是這樣: 大部分人都是普普通通的。

普通或特殊是在和類似的人相比之後的結果: 多半的人一切都普普通通。因爲那就是普通的定義, 你若是在前面的百分之三, 那你當然不那麼普通了, 你若是在後面的百分之三, 那你也當然不那麼普通了, 但94% 的人既不在前面三個百分點, 也不在後面的三個百分點裡。百分之九十四的意思再明白也不過, 在街上隨便找一百個人, 平均有九十四個是普普通通的。

這裡面有學問嗎? 這樣的簡單道理, 粗粗一看, 是沒有甚麼學問的, 但用一點點畫圖的技術, 問題馬上就不一樣了。當然, 也就有了前、後的百分之三以及當中的百分之九十四。圖一是民國 77 年度大專聯考裡兩萬多名考生的英文成績。圖一所畫的是所謂最簡單的長條圖。

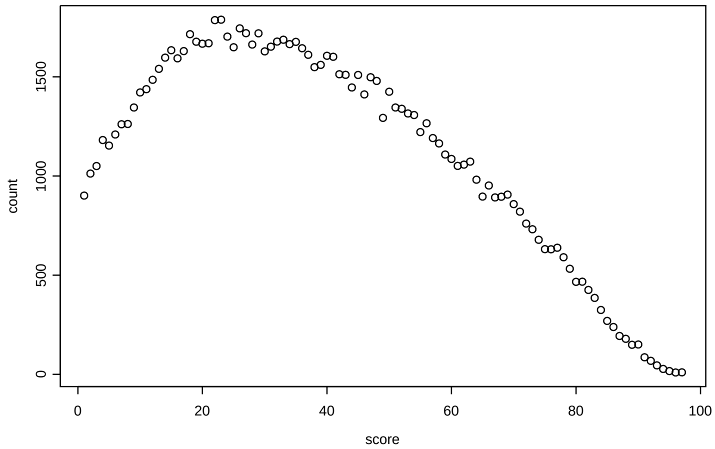

Figure 1: 107272 份英文成績的次分布

。

這個圖像甚麼? 這差不多是但又不全是常態分布的曲線。這當然不那麼像常態分布, 但也差不到那裡去, 反正它基本的道理是十分清楚: 極差和極好的考生之外, 都是普通考生。普通者, 佔大部分也。

計算一下這兩萬多學生的平均值和標準差。圖一裡的曲線是利用這兩個值所算出來的常態分布曲線: 是相差不多是罷!

我們輕輕鬆鬆地, 幾乎是「芝麻開門」一般地打開了一個重要的寶庫: 常態分布。

你在中學有沒有要證明根號二是無理數?

你有沒有在微積分裡學過 $sin(x)/x$ 在 x 趨近於 0 時的極限值是 1 , 而這裡的「正弦函數」得要用「弧度制」, 而弧度制, 得靠圓周率等於 3.14159... 才能定義出來?

你有沒有學過複利, 而用到 $e = 2.71828\ldots$

如果你覺得這三件事都有一點學問的話, 那麼常態分布的學問可能更大, 它的標準型 (意思是說最簡單的) 公式是

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^3\} \tag{1}$$

我們既用到根號二, 也用到圓周率, 還用到自然對數的底 e 。把任何一組數目, 如果它們有一個共同的題目, 如果它們由測量—— 不論是自然的方式 (如量你的腰圍), 還是人爲的硬性規定 (如考試的分數在 0 與 100 之間),—— 如果將這一組數字畫一個長

條圖, 那麼圖一的形狀, 便很有可能出來。

有甚麼用? 中國人做學問, 基本上上焉者是爲了經世濟民, 下焉者爲了自己吃飯。爲好奇而做的學問就少了。中國人有興趣的是用處。但純粹自用處的觀點來找重要的結果, 是很難的。重要的發明或發現, 原動力來自「爲甚麼」, 而非「有甚麼用處」。一大堆純粹因爲好奇而導出的事物也許都是沒有用處的。但偶然也有一些極有用的, 常態分布是其中之一。

一個顆刻磨成58面的鑽石有用嗎? 它可以切割玻璃不錯, 它也可以討好美女, 但它的價値來自它的完美: 硬度無出其右, 折光率也無出其右。何況研磨的技術是人類能力的尖端所能, 它的價値來自它的性質。它是自然生成的, 在極難的條件下生成。

對統計學者而言, 常態分布的存在就如鑽石一般: 它不是我們發明的, 它是自然產生的, 而人類用極大的心力從自然現象裡逐步提粹精鍊, 而最後得到前面的公式 (1) 這樣的簡潔形式: 有如一座皇冠, 鑲上了e及圓周率兩粒閃亮的寶石, 而尙嫌不足, 又配了一個用根號二做的鍊子。

但這是只是形而上式的讚美, 常態分布的精要, 來自它所具有的性質: 是這些性質讓這個分布成爲一個重要的分布, 而不是任何的人爲因素。例如它又叫做高斯 (Gauss) 分布, 但並不是因爲高斯導出它它才偉大 (事實上高斯的推導, 還有一些繞圈子的邏輯未能釐淸), 它的重要來自它本身的光輝。

大家都知道的一條性質是:「常態分布加常態分布, 仍然爲常態分布。」這經常是標準統計或機率教本裡一定會講到的性質。我們現在當然不用功夫討論中央極限定理, 但是這一條性質以及它的逆定理:「如果有兩種分布的和恰好是常態分布, 則這兩種分布都非同時是常態分布不可。」這兩項重要的性質, 左右夾擊, 使得中央極限定理不止是非成立不可, 並且它極限分布, 也非得是常態不可。

我們先自可看到的一組一組的數據, 來看是不是這樣有的分布, 然後再經由一些整套的數學推導 (以及近200年一大票學者的努力), 其結論是: 我們所能收集的一大堆數據裡來看, 如果我們鐵口直斷地說它們是來自某一個常態分布的, 那麼, 我們不會太錯。

這個道理的重要就和生物學家要懂得細胞結構, 物理學家要知道原子結構是相差不多的。常態分布是我們所最了解的數據的基本結構。而我們用它來描繪人類最最不能避免的問題: 誤差及變異。變異所代表的意思是同樣的物件或事情, 當它一再重覆發生的時候, 不論我們如何用心測量, 總是不能得到同樣的結果, 例如生產線上的規格, 絕不可能全然一絲不少地合格。再用心挑選的儀隊成員, 其身高體重都不會一樣, 只是有的變異大些, 有些變異小些, 而變異在常態分布的時候, 可以容易地用一個尺度 σ 來表示:

$$\phi(x;\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}x^2\} \tag{2}$$

和前面的公式 (1) 相比, 這相當於公式 (2) 裡令 $\sigma = 1$ 的特殊情形。我們知道 σ 的值愈小, 則變異愈小, σ 的值愈大, 則變異愈大。圖二所示的是 $\sigma = 0.5, 1$ 及 2 的三個情形。

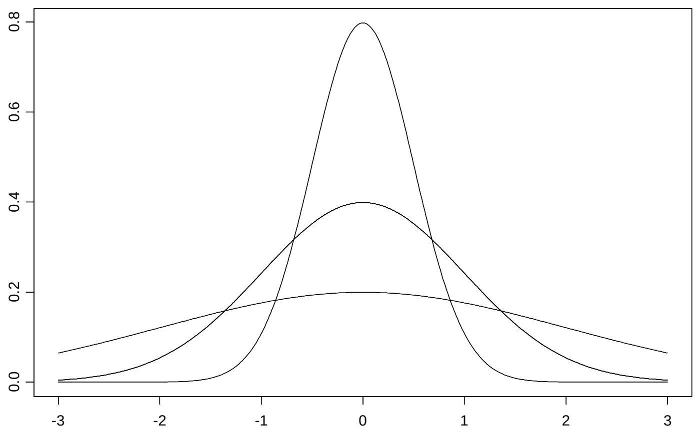

Figure 2: 三個不同的標準差所產生的差異, $\sigma = 0.5, 1, 2$ 。

我們當然也可以將誤差想像成變異, 但我們還可以更多想一想。自從牛頓發明了第一到第三定律, 我們就比較願意去想像宇宙萬事, 不論是自然的, 還是人文的, 大概都有一個方程式可以表現出來。

這當然是一件了不起的想法, 我們只要想一下力學的能憑幾個公式就計算出多少道理, 以由此而產生的近代科技文明, 只要有一點想法的學者甚至政客, 都會不禁地想: 假如其他的事情也可以用公式來表示, 那會有多厲害!

但世事豈能如此順心!我們才學懂高中物理的時候, 多半以爲世上的事情都可以用公式計算, 等到年紀愈大, 見得愈多, 才知道人類眞正能夠計算的東西, 遠比他不會計算的東西爲少。但這並擋不進我們想去「抓住」一些現象並寄望由此能做預測的想法。通常, 我們能抓住, 或者至少希望能抓住的那一部分, 叫做模型, 而抓不住的部分, 叫做誤差, 而

$$見到的現象 = 模型 + 誤差$$

這裡, 誤差若是用常態分布來表現 —— 其實也不會非常離譜 ——我們就差不多概括了目前人類知道的大多數的計量模型。我們當然是將問題極度簡化, 才有以上的敍述。但精神上這差不多就是目前大多數計量方法的精義。這些得過諾貝爾獎的, 還沒有得但非常想得這個獎的學者, 過去一兩百年甚至未來一二百年, 多多少少地用這樣的基本想法來利用資料與模型, 既歸納又演繹地尋找人間萬事的規則。這些工作, 若沒有常態分布, 免談。

人們之所以能夠這樣做, 一個基本的道理是我們已經非常了解常態分布, 以至於我們可以充分地透過它的良好性質, 一方面安心地設立模型, 另一方面由這樣的模型裡做出推論: 我們對大部分天下事物的了解, 並不是由於我們眞的懂得其中的運行規則, 倒反而是因爲我們比較了解問題中的變異及誤差!

早年的農業實驗, 希望能找出最佳的肥料比例以使得稻米的產量更高。這一系列的活動, 背後都是基於我們對常態分布的了解而達成。至於二次大戰以後就一直不曾停止的工業品質與工程的工作, 也一直在常態分布的支撑之下有成。農業的改良讓我們吃飽, 工業讓我們的衣、住和行, 甚至育樂得到滿足。那麼, 新藥的發展 ——爲甚麼它們能夠既有效又無害? 爲甚麼我們對它的效應及害處都能夠用計量的方法加以詮註: 這也得力於我們對常態分布的了解。至於人文及經濟上的活動更不必說了。幾乎可以說, 沒有常態分布, 而這些學問都只能停留在原始的性向描述的階段而已, 那裡談得上任何解析的結果。

可是, 這些模型也太沉重了一點。我們不必那樣在意國民所得, 你說, 那是主計處的事。我們也不必那樣在意考題的信度和效度, 你又說, 那是教育部的事。如果常態分布只是用來做這些官方問題, 那麼我不知道也罷。

那麼你知不知道第四臺有一位老師, 號稱是留美的碩士, 並且用迴歸分析「這樣高級的統計方法」來解盤?

這我其實所知也是有限。但是以常態分布爲基礎的統計模型, 的的確確在證券分析市場裡大賣特賣。其中最有名的一個模型, 叫做「美國 BARRA 多因子淨値模型」。它自1970開始有了基本理論基礎以來, 目前已使其公司達到年營業額四千萬美元以上。至少600家主要的投資公司 (它的總資產額超過一兆美元) 長期訂購這個模型。其目前所賣的 E2模型, 卻全然是傳統的統計建模 (modeling building) 工作。

這樣的工作有多難呢? 在技術面上, 並沒有想像的那樣難。如果有源源不斷而來的

定期數據, 如果我手上有一個像樣的主計處裡面一個技術科的人力, 努力兩三年就應該有可以出售的商品了。事實上,BARRA 模型裡用了不下60個因子, 算是一個相當複雜的多因子模型。如果只有一個因子, 即所謂的單因子模型, 那麼各證券的投資報酬率針對那個唯一因子 (多半指某一種市場指標, 例如我國的發行量加權指數) 的敏感度, 就是目下較高級一點的證券分析師口中所說的「貝它」(beta) 值。貝它有甚麼用? 它可用來測量某一證券針對某一證券組合的影響。還太難了是不是? 且不必管這些技術名詞, 更且不必管這些名詞的眞正意義。在投資工業裡, 貝它值是常年地被各家各派的高手計算以及估計著的量。「貝它服務」是可以像訂報紙一樣的按時送達的。

整個投資分析業是大量用到常態分布的行業。只是傳統的統計學者, 都太帶著理學院的學者氣息, 不願意眞正去碰這個行業, 因此, 表面上它的統計意味不那樣重。但在這裡, 無數個聰明精巧的心靈, 極盡所能地從如海的 data 中, 希望比他人更早一步找出重要的規則與信息。在另一方面, 一個有效率的市場, 其定義卻正是「應該無法自公衆皆可得到的 data 中找出必然賺錢的方法」的市場。這是訊號和雜訊交戰的殺場, 也是一個近乎零和遊戲中永不休止的震盪與徘迴。這是一個隨機的世界, 而隨機乃是常態。

人類因爲有數字的概念而比其他動物高明。我們曾經以爲這方面的智識是絕對的:「給我夠長的槓桿, 我可以舉起地球!」但現代的人類要謙卑多了, 他們知道多半的時候, 力有未逮。而在他們承認了誤差及變異的存在之後, 反而海闊天空地走入另一片開廣的領域。萬物有常, 這是我們想抓住的部分: 世事多變, 這是我們用適當地常態分布來表現的部分。這個道理, 好像易經裡也有。但是, 將常及變放在一起, 加以計算及演繹, 並由此做科學的驗證, 卻純然是近五六十年統計學及相關計量科學和廣用科學的發展。

我們都是普通人, 這是「常」的部分。普通人也有高有短, 有胖有瘦, 這是變的部分。這些工作, 只是儘量抓住「常」的部分, 而讓「變」的部分受到控制。而這一切都希望用眞實的數據和確切的計算來支持我們的論點。我們當然不只是談身高體重, 我們談國民所得, 失業率, 國家競爭力, 品質工程, 大專甄試的信度和效度; 衛星信號的濾波, 由報稅單裡看出異常而加以防止逃漏, 環境空污—— 這些問題的裡面, 都離不了數據, 模型及誤差。而沒有我們對常態分布的了解, 要想有效地處理這一類的問題, 門都沒有。

常態無所不在。因爲這就是「常」的意思。月有陰晴圓缺, 但因爲我們眞正了解到沒有嫦娥玉兎和吳剛, 我們反而不會失落。分子與分子的碰撞該如何計算? 因爲分子實在太多, 我們實在無法用撞球臺上的物理學來描繪分子碰撞的軌跡。因爲我們知道那樣做不行, 所以我們反過來認定我們無法全然知道碰撞後的路徑, 用隨機的角度來想。這似乎是破釜沉舟的法子。但這也是愛因斯坦用以算出布朗運動公式的法子。

人生因爲有兩個因素而變得瑰麗: 可以努力改善但又不能預測。人生又因爲兩個因素而變得無奈: 無法溝通而又需要溝通。這些都因爲一個基本道理: 世上並無完美, 而智慧只來自對缺陷的體認。這個世界是如此地糟, 因爲一半的人都是混蛋; 它又如此美麗, 因爲一半的人都是君子。用我們的話來說, 那就是說: 如果 X 是常態, 則 $-$X 也是。

讀到這裡, 你應該可以懂得, 爲甚麼太極圖有兩種讀法: 白魚有黑眼睛, 黑魚有白眼睛。但要這個道理用實在的言語、公式或者計算加以闡述, 而讓河圖、洛書、易經這些東方的古早智慧不流於風水、命相、五行、八卦之學, 從而能漂漂亮亮地對世事萬物的現象, 在能解釋的時候解釋, 在不知道的時候也知道並承認尚不能解釋, 然後我們才能從玄學中昇華出來。

統計學是我們適當的搭配了我們對常態分布的了解 (也就是說對於誤差及變異的了解) 之後所衍生出來的歸納和演繹的方法。用統計建模的方法解釋世事萬物的現象, 除了有「知之爲知之, 不知爲不知」的誠懇之外, 因爲再加上對於不知部分的科學描述, 而增加了分析的力道。而這在以前, 是完全沒有辦法做的。

可以這樣說: 發現並發展一套以常態分布爲基石的分析方法, 是二十世紀人類的重大成就。這套統計建模的方法眞的有用嗎? 我們用一些方法, 並不一定知道它們是對的。有很大的一部分理由, 反而是除此之外, 並無別法。造物主是何等神奇, 人類的知識, 又是何等的淺薄無助。這一點點微光, 在自然的神奇、甚至在人文的多變之下, 既是極力的掙扎圖存, 又是非常的力不從心。因此眞正的智慧, 反而來自謙卑:「所有的模型都是錯的, 但有些是有用的。」說這樣話的人, 應該是眞正的智者。

file=D:/ My Stat Article/paper10.ctx, printed February 7, 2007