知識天地

多維空間上樣本點的大小與隨機性

黃顯貴特聘研究員、蔡宗希副研究員(統計科學研究所)

日常生活中可能常會遇到類似這樣的選擇問題：買350cc小罐裝的鮮奶每罐28元，大罐一公升裝每罐75元。簡單的常識告訴我們，如果品質一樣，當然買“俗擱大碗”的選項。但話說回來，本能上也許會質疑便宜貨的食安問題。雖然買多量單價卻較高的烏龍事，偶爾也會發生，但一般而言，大罐的應該“卡俗”。所以我們計算一下，平均每一元，小罐裝可買12.5cc牛奶，而大罐裝的則可買到13.33cc，果然是大罐裝的比較便宜一點。

如果將買鮮奶的例子轉換成數學語言來敘述，鮮奶的量與價格可以看成二維的數值：(350,－28)與(1000,－75)。這裡用負號，因為買東西是付款而非賺錢。我們知道多維度的值常無法立即做出大小比較。但如果轉換成一維的數值：12.5與13.33，則答案便很明顯。上述的例子恰好有客觀標準(即比較它們每一單價可購買量)可轉換成一維數值，但是其他例子不見得適合這個做法。例如，某次學校考試成績如下(參閱圖一左)：

小明：數學94分、英文85分。

小華：數學88分、英文93分。

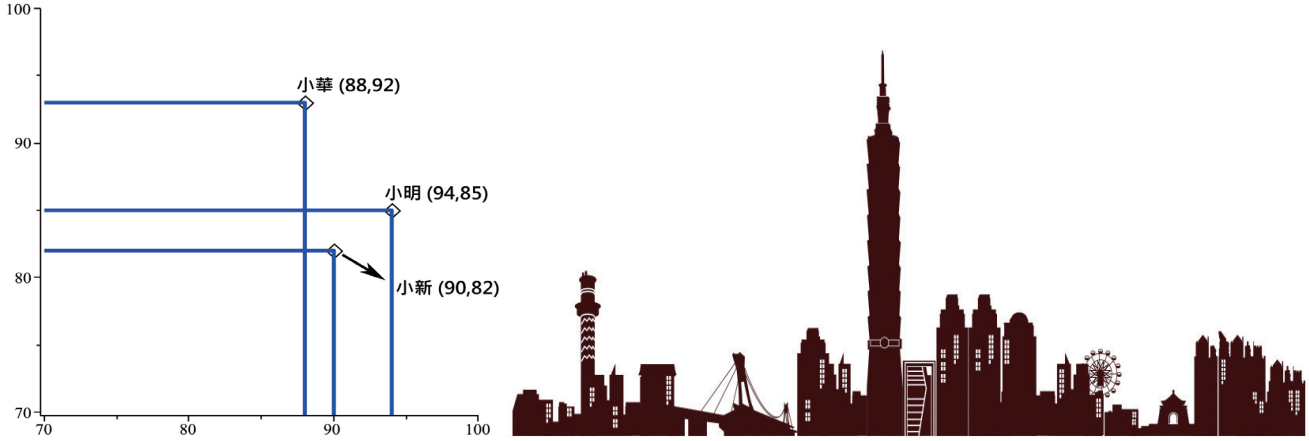
小新：數學90分、英文82分。

最常用的排序法是看總分：則小華181分>小明179分>小新172分。也有些方式是看加權後的總和，例如若是數學加重50%，則名次變為：小明226分>小華225分>小新217分。不同的加權比重可能產生不同的名次排序。但是我們發現，不管採用任何一種加權比重，小明都優於小新，因為小明數學與英文兩科的成績都高於小新。

上述小明(94,85)優於小新(90,82)的關係，在理論上可以作如下的推廣。取d維度空間中的兩點A與B，如果點A的每一分量座標值皆大於或等於點B相對應的每一個座標值，則我們說點A壓制(dominate)點B。這個簡單的壓制關係是多維空間中，點對點關係最直觀也最自然的偏序(partial order)，它也在極多領域被廣為運用。從統計上的admissibility、工程上的elitism、計量經濟上的efficiency與Pareto optimality，到計算機上的maxima與skyline，離散機率上的sink或source等，皆源自同一壓制概念。日常生活上也不乏利用此概念的例子。

有了這個壓制關係來衡量點與點之間的相互順序後，我們可以問，一個樣本中最大的點是哪些？這與一維的情況不同，因為並非任兩點都存在這種壓制關係(如上例的小明與小華，無法比出大小)。如果我們定義最大的點是壓制其他所有樣本點的點，則滿足這個條件的點可能並不存在。所以，比較合適的定義應該是：“最大的點是沒有被樣本中的其他點壓制到的點”。

以上述考試成績的例子來說，小明與小華都是此三人樣本中最大的點。給定一個樣本，我們稱上述定義得出來最大的點為『極大點』。近年來許多文獻，尤其是計算機上的資料庫語言，把這些極大點的集合統稱為『天際線』(skyline)。這個名稱相當符合幾何視覺的直觀(如圖一右)，以下我們採用天際線這個名稱。



圖一　三人分數的相對位置(左)與台北的天際線(右)。

我們開始介紹天際線這個統計量理論上的結果。在維度d的空間上，給N個隨機點，並且假設這些隨機點在各維度的分量數值是獨立分佈。文獻上，Barndorff-Nielsen 與Sobel(1966，文獻1)最早研究天際線的機率分佈性質。他們導出如下的期望值表達式

$$M_{N,d} := \sum_{k=1}^{N} \binom{N}{k} \frac{(-1)^{k-1}}{k^{d-1}}$$

並且證明當N很大時，它近似於

$$M_{N,d} \sim \frac{(\log N)^{d-1}}{(d-1)!}$$

從數學的宏觀來看，這期望值$M_{N,d}$有很多種不同形式的數學表達式，有遞推、機率、有限和及積分等面相，也反映出問題結構的多元與豐富性。

班氏與索氏同時也在1966年同一篇論文中考慮天際線的變異數的行為，他們導出d=2,3時的表達式以及漸近分析的結果。而這問題的複雜度隨著維度增加變得困難。到了1998年，我們(文獻2)證明天際線的變異數與其期望值漸近上是同階的(即大概近似於$(\log N)^{d-1}$的倍數)，我們也進一步導出漸近展開式的首項係數表達式。而更精確漸近分析結果，在後來幾年也陸續被釐清(文獻3)。

我們知道期望值、變異數以及極限定理是研究機率最重要的三個性質。我們在2005年完整地證明天際線的中央極限定理，亦即天際線的個數當樣本點增大時趨近常態分布。古典的中央極限定理的發現可以說是機率學發展史上重要的里程碑。這定律最簡單的產生方式，便是丟擲N個銅板並且紀錄正面出現的次數。這樣連續做M次，當N與M皆夠大時，正面出現次數便近似於常態分佈。另一方面，每次丟擲銅板時，其結果的獨立性是中央極限定理成立的重要關鍵。推廣到一般的狀況，一個統計量如果可以被看成數量很多個獨立隨機小變數的總和，它應該會漸近到一個常態分佈。

例如，在適當的獨立隨機性假設下，考試成績的平均值也會漸近常態分佈。因為平均值為總和除以學生總數，所以漸近常態分佈是容易被導出的結果。而天際線的情況就不明顯了，需要深入了解它隱含的內在結構，找到適當的變數轉換，然後切割出大部分相互獨立的隨機小變數，並且驗證剩餘相關的小部分在漸近上是可以忽略的。上述的方法也是證明中央極限定理常採用的策略。

上述的理論結果是在各維度的分量獨立性的假設下成立。如果各維度的分量存在相關性的話，天際線數量會大不相同。簡要地說，如果各維度的分量是正相關，則天際線的數量是O(1)，也就是不管N多大天際線都有界。反之，如果各維度的分量是負相關，則天際線是以$N^{1-1/d}$階的速率增加。這方面文獻上主要的討論是集中在d=2，在明確的隨機模型下，可以導出精確的漸近分析結果。

我們總結天際線的優缺點：它的優點是可以簡單快速篩選出相對優勢的資料。而缺點是，天際線的大小是“先天”注定的，不一定符合實際應用的需要。當天際線太大時，則失去篩選的意義。反之若天際線太小時，恐有重要資訊遺漏的顧慮。再加上資料在取得時可能已經隱含了一些誤差，原本的優勢點可能因此被“誤殺”，而產生遺珠之憾。

天際線太大，對於高維度樣本幾乎是無可避免的問題，這是所謂『維度詛咒』現象。如果太多的點是天際線，這樣的天際線沒有多大用處。近年來的有很多研究考慮將天際線精緻化。其中數學上比較簡明的是偏壓制關係(k-dominance)的研究。其定義如下：在維度d的空間中的兩點A與B，如果點A存在有k個座標皆大於或等於點B相對應的每一個分量，則我們說點A偏壓制點B。當k=d時，偏壓制關係就回到原始的壓制關係。

偏壓制關係是全壓制關係很自然的推廣。但是，偏壓制關係缺了遞移性，也就是說當A>B與B>C成立，卻不保證A> C成立。因此可能會有循環出現。例如，當d=3, k=2時，(1,2,3) > (3,1,2) > (2,3,1) > (1,2,3) 。於是它最大的缺點便是由它定義出的天際線可能會不存在，這樣的現象，在固定維度的隨機樣本點上，卻是無可避免。較精確

一點來說，我們證明(文獻4)如果維度固定，樣本點數量增加，天際線期望值趨近於零。這意味著實際上使用這類關係時，必須更加小心。而比較有趣的模型是讓維度d也隨著樣本大小N來變動，這也比較符合實際使用的狀況，因為實用上無限大並不存在。這種模型下，所有的分析問題難度劇增，因為必須同時考慮d與N同時變大的均勻估計。極為有趣的是底下的閾值現象(文獻4)：當d小於一個臨界值D時，天際線的期望值趨近於零。另一方向，當d大於D+1時，天際線的期望值趨近無限大。而當d=D時，這個值在0與一個常數C間變動。最後當d=D+1時，則在C與無限大間震動。這裡，D約在 $\sqrt{\frac{2\log N}{\log_{\log\log N}}}+1$ 而 $c=\frac{e^{-\gamma}}{2-e^{-e^{-1}}}$ (γ 是尤拉常數~0.57721)。這樣漂亮且精確的結果，在文獻上並不多見。

不管是閾值現象，或是常見的相變現象，都可看成是以一種方式來描述兩種或多種現象，在結構或數學上都有極高的興趣與研究價值。典型的問題包括：結構上為何產生變化？在哪裡產生？如何變化？一般性在哪裡？在數學上如何描述？需要什麼工具？與哪些問題相關？微觀是否有更深層的變化？從一個現象出發，是否能開發出整套理論(包括新的方法)，進而對實際問題有所助益？

隨機樣本在許多科學領域上，除了扮演重要的理論與實務間橋接的角色外，亦經常引出許多有趣的現象與極具挑戰性的數學問題，同時如何在模型的實用合理性與數學上可被釐清間取得平衡，皆是相關研究引人入勝之處。

參考文獻

1. O. Barndorff-Nielsen and M. Sobel, Theor. Probability Appl. 11 (1966), 249-269

2. Z.-D. Bai, C.-C. Chao, H.-K. Hwang and W.-Q. Liang, Ann. Appl. Probab. 8 (1998), 886-895.

3. Z.-D. Bai, L. Devroye, H.-K. Hwang and T.-H. Tsai, Random Structures Algorithms 27 (2005), 290-309.

4. H.-K. Hwang, T.-H. Tsai and W.-M. Chen, SIAM J. Comput. 42 (2013), 405-441.