# 知識天地

## 自在寬容的學習和使用統計　探索與精緻學習

張源俊研究員（統計科學研究所）

統計難，難在它的數學語言，也難在它的計算工具，更難在它貼近實務。這樣說來，成為一個好統計學家似乎得有三頭六臂。 曾經有前輩統計學家對我說「統計就是服務其他學門」；也有數學系的前輩老師對我說，他們系上學統計的學生沒法處理實際問題是因為「數學不好，抽象化的能力不夠」；然而在「巨量資料」的潮流下，我的同事們卻說現代的學生學不好統計是「沒有data sense」，加上寫程式能力太弱之故；但卻也有人說：「統計本身的問題都做不完了，哪有空去管機器學習或其他資訊相關的統計」。我個人認為這就像機率學者鍾開萊（Chung, Kai Lai）在他的書的序言中寫的：「One man's technicality is another's professionalism」。

在2002 年，MIT有一篇標題很特別的博士論文叫做「Everything old is new again」，主要談論因為「機器學習」被重新重視的一些數值最佳化算則。從統計學家的觀點來看，這種情況會鼓勵和刺激大量新的研究和開發更多的工具，特別是那些因採用進階或複雜計算優勢而變得實用的「密集計算型統計方法」。然而我們必須知道，統計不僅是分析工具或算則的集合。為了自在寬容地學習統計，自在寬容地學習從數據中獲得有用的訊息，並將其用於我們的工作中，我們更需要了解統計的核心 — 所有方法和算法背後的核心思想。當然我不可能在這短文涵蓋所有的故事，所以我將以一些古典統計論文為例，引起大家對這些重要的古典方法和理論背後的原始想法的興趣。依個人興趣，我選了 Fisher（Sir Ronald Aylmer Fisher，February 17, 1890 - July 29, 1962）、Wald（Abraham Wald, October 31, 1902 – December 13, 1950）和Tukey（John Wilder Tukey, June 16, 1915 – July 26, 2000）等人三篇二十世紀早期的論文。以這幾篇經典論文為例，作一簡單的說明，我們將體會到這些想法如何與現今「資料科學」的連結，並進而認識到好的「核心」想法是可以活躍於各個世代並延續下去。

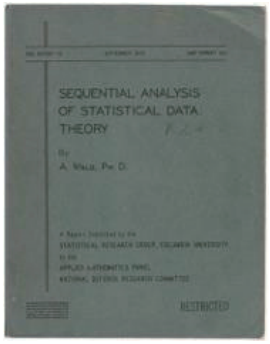
Fisher在「統計的數學基礎」論文中，引入「充分統計量」的新概念，並以此說明為什麼數學知識對統計學至關重要。為此，他強調了數據分佈的特性，並且在分佈假設下去討論充分統計量的性質。自那篇文章發表以來，許多分佈的性質都被發掘出來。依循分配的屬性，他成功地引入數據「簡化」的方法（data deduction）— 稱此概念為「充分統計量」;也就是說，如果您願意或能夠做出一些分配假設，那麼您不必隨時攜帶所有數據跟夥伴討論，簡單記錄充分統計量的資訊就夠了。因為您可以藉分配假設和其相應的充分統計量來重建整個隨機結構。由於他的成功，奠定往後數理統計學發展的基本方向，後續的研究僅是讓數學方面的想法或假設更為深入。

眾所周知，「假設」相當於「限制」。每次你對自己的問題做出一個假設，就會為自己的問題劃下界線 — 定義了一個「舒適區」。也許其中一些假設是來自於問題本生，且有針對問題；然而有些假設只是讓我們能方便使用既有工具，或僅僅是因為我們無法在這些假設外，建立很好的理論支持。這些年來，我們已從許多經驗中知道，即使有時假設沒有實現或確認，在許多實際情況下，結果「似乎很好」。相對地，我們也經歷了一些重現性（reproducibility）出問題的狀況；雖然在某些問題情境中能得到不錯的數據結果，但在其他實際情況下，成功經驗卻不能重複，這對於科學研究的信賴度將造成很大的折損。另一方面，也由於現代數據的複雜程度，造成我們這個時代的人們想拋棄分配假設，進而也開始質疑Fisher 的充分統計量概念在現今數據科學時代是否適用。

如果我們把自己放在他的時代，當沒有太多分析數據的計算工具可用，而我們仍必須與他人交流數據信息時，最「誠實」的方法是什麼呢？困難並不僅在於計算，還包括如何呈現資訊。想用數據可視化嗎？別忘了，現代計算技術可是在那論文幾十年後的產物。如果我們能夠以Fisher的觀點來看問題，或了解到 Fisher的「充

分統計量」是可以被視為「數據簡化」（data deduction）的工具，那麼從他的論文引出的後續發展將會有很大的不同。「分配假設」將只是用來定義一個「舒適區域」的數學語言，讓他可以使用手中的數學工具來發展這個概念。我們也將進而意識到，他這篇出名的論文，最重要的概念其實就是簡化數據，而其自然的延伸將是「如何在沒有分配假設的情況下簡化數據？」。如果有辦法這樣做，那麼由數據大小建立的計算障礙就可以很容易去除。

「在沒有分配假設的情況下，是否可以簡化數據，同時保留充份統計的想法或概念？ 如果沒有，那麼最有希望的近似方法是什麼？」 以這種方式思考問題，那麼統計科學的發展與過去的歷史將會有很大的不同。我們如何能夠發展出一些可以充分利用現代計算機和計算能力的新穎統計概念或方法。例如，如果數據存儲在有利於排序結構的計算機中，我們將可以忽略計算平均值和中位數的計算時間上的差異。現今的計算技術中，應該有越來越多這樣的例子。我們如何在不同數據結構中，找到一個有益於計算密集型方法的數據儲存結構（資料結構）將是一個有趣的研究課題。如此才是掌握「統計核心」，也才能自在寬容地學習統計，並在用統計方法時也能自在寬容，不侷限於複雜的技術性假設。




如同許多學門，統計科學也有很多面向。在數據分析過程中，統計是很有價值的工具。統計人員通常透過與領域專家的合作，實際了解資料收集的流程或資料收集時的操作狀況，並在維持「統計核心想法」下，給一些特定的應用情境設計或開發合適的方法。如果你仔細觀察，統計核心價值和其理念其實是一樣的。在第二次世界大戰期間，Wald開展的「次序概率比檢定」（sequential probability ratio test）就是一個典型的例子。在這段戰爭時期，人們試圖更快地檢測製造武器是否良好，以確保武器的可靠性。事實上，這篇文章一開始僅允許在某些統計學家之間流傳，為國家利益不准外傳 ，所以在論文報告封面上可以看到restricted的字樣（參閱左圖）。Wald 在該文件所表達的想法其實很簡單，他允許根據「隨機數據樣本」來決定樣本數大小。這與當時的主流統計很不一樣，反而跟我們日常的決策行為很像。這又一次告訴我們，很多假設僅是讓我們找到「舒適圈」，並非為了徹底解決問題。Wald 這一個隨機樣本的想法，後來被進一步發展應用於臨床試驗上，以縮短是否通過新藥或醫學方法的決策時程，使好的藥或方法可以盡快用於病人身上。這在醫學相關研究中仍然是流行和重要的議題。除此之外，Wald也啟發許多研究者開發新的方法。譬如，當讓樣本是依據特定條件在可控制的情況下被引入分析時，我們如何分析這「非獨立樣本」的資料。這想法與目前的「精準醫學」息息相關，甚至與「主動式學習」有密不可分的關聯。Robbins和Monro（1952）的隨機逼近就是一個例子，他們的問題的數學定義明確且工程中的優化問題通常就是這種類型的 — 當您沒有一個清楚的數學模型（或模型太複雜），但仍然希望能通過含有雜訊的觀測值來估計，透過調整某些參數來進行優化。這篇論文被引用了非常多次，而這個想法後來被廣泛應用於許多領域，包括現代的教育測試、臨床試驗中的適應性實驗設計，甚至於精準醫學。這概念也適用於現代「推薦系統」及其他類似的應用。 亦即利用系統中既有資料的資訊去收集新的資料（sequentially and self-adaptively data recruiting processes），並且基於這些隨機觀測樣本，構建一些預測模型。

每當有新形式的數據 — 包括數據格式，數據收集，大小等等的改變，人們就會再次質疑統計方法的可用性。歷史上，人們曾多次詢問過這個問題。因為現有的方法是在不同情況下，針對既有問題開發的。然而別忘了，其核心不變。能掌握統計核心就能自由寬容地使用這些已知的工具。 也許因為數據運用的環境的不同，我們需要不同的操作方式、但從數據中挖掘出有用的信息，應該是所有從事「數據科學」的共同目標。而這些信息將是我們制定決策的依據。由於錯誤判斷的代價有大有小，對於一些問題，我們可能需要更多詳細的信息，與此同時解釋的能力也將會提高。對於這些問題，需要一個有效學習數據分析的程序。探索與精

緻學習各有其角色。科學研究一向都是「踏著前人的肩膀」往上，探索式的學習讓研究者能快速從資料中汲取訊息，提供了更多想像空間，以提出進一步的假說；精緻學習則幫助我們從各個層面進一步的確認，讓訊息成為可累積的知識。

知識的累積和教育的目的就是要讓一些原本是廟堂裡的知識，成為普羅大眾的常識。而非像「巫醫」一般，把「知識」視為操弄無知群眾的工具。統計學為了服務其他學門，對學生的訓練必須隨時代而改變，而一些統計學上的「知識」和技術也該成為其他行業的「常識」。當其他人都在用我們前輩們所發展出的方法和概念去處理問題時，我們除了高興統計有用，也該想想如何讓統計在下一波的知識「革命」中可以更有貢獻。課程的改革是關鍵。如果只是關注眼前的時尚潮流，那我們就僅是學術圈的「啃老族」罷了。任何學門的存在必需有其核心價值，統計的核心在哪，透過技術的學習我們也希望傳承統計的核心思想。有位前輩曾說統計學的基本就是mean 和 variation，這說法看似簡單，但衍生出的概念卻很廣。所有統計方法都希望讓我們把資料看得更清楚，然而如何「看」卻是難題。mean 和 variation 是看資料的基本，mean 想表達的是「集中」的位置，或許兼有「代表」的意涵，而variation想表達的是「分佈」；當這兩個量不夠時，我們會用些假設來簡化問題，就此目的而言，機率分布的概念是個很好的數學工具。

資訊科學的情況和統計很相似。所有人都在用電腦，他們做了很多服務各行各業的事，也為了能「服務」得更好，他們想要有更好的「計算」方式，從各種軟硬體的層面來達到此目標。所有領域都有寫程式的需求，但不會每個人都在寫「作業系統」或使用「組合語言」。*Software tools*（Kernaghan 和 Plaguer）這書出版於1976年，迄今已四十年，這書堪稱程式設計的經典著作；書上主要的概念在提醒寫程式的人不要寫那種「可拋棄式」的程式 — 僅僅為了特殊目的，用完之後就沒有價值。他希望教大家寫可以重複使用到類似應用情況的「軟體工具」。在發展或研究統計方法時也一樣；用tuning parameter這概念為例好了，現在很多方法都涉及一些參數的選擇，那麼所發展出的方法到底是一次性、可拋棄式的，還是一個工具，在此類涉及參數調整的問題中最容易分辨了。

回歸「想看清資料」的初衷，去尋找合適的工具，可能才是統計本質。然而工欲善其事，必先利其器。加上統計學的研究，不僅是工具上的限制，還得考慮資料本身的限制。對工具的了解是選擇工具的基本。數學是工具，計算機程式是另一種工具；了解問題所涉及的內容，則有助於選擇適當的工具。懂得越多，選擇越廣，也越精緻，成果自然也不同。

Tukey 有一篇非常著名的論文 —〈數據分析的未來〉，這當年可是發表在 *Annals of Mathematical Statistics* 的文章。 在Tukey 的論文中，他已經討論了數據導向和無分配假設的方法的概念，可見這些也不是新的概念。他也提到了有關數據「可視化」的一些想法，還強調「數據可能不包含答案」，也指出「數據的組合和對答案的渴望不能確保可以從給定的數據體系中提取合理的答案」。他更強調 「我們應該找出全新的問題來回答」；「我們需要在更現實的框架中解決舊問題」；「我們應該找出不熟悉的觀察資料摘要，並確定其有用的屬性」（We should seek out unfamiliar summaries of observational material, and establish their useful properties.）等等。在他的論文中，他還提到了一些關於「最優化方法可能產生的危險性」，而每個人都知道這正是我們現代數據科學所面臨的問題。此外，我們需要能「問對問題」，「正確的選擇方法/技術」，「正確的選擇標準或測度來比較結果」。最後，我們以Tukey的著名引語結束：「對於一個正確的問題的近似解好過於對於一個錯誤的問題的確切答案」（It is better to solve the right problem approximately than to solve the wrong problem exactly.）這句話總是準確的。