數據分析那些事

8.2K Followers About Follow

統計學權威盤點過去50年最重要的統計學思想，因果推理、bootstrap等上榜

數據分析那些事 Dec 3 · 27 min read

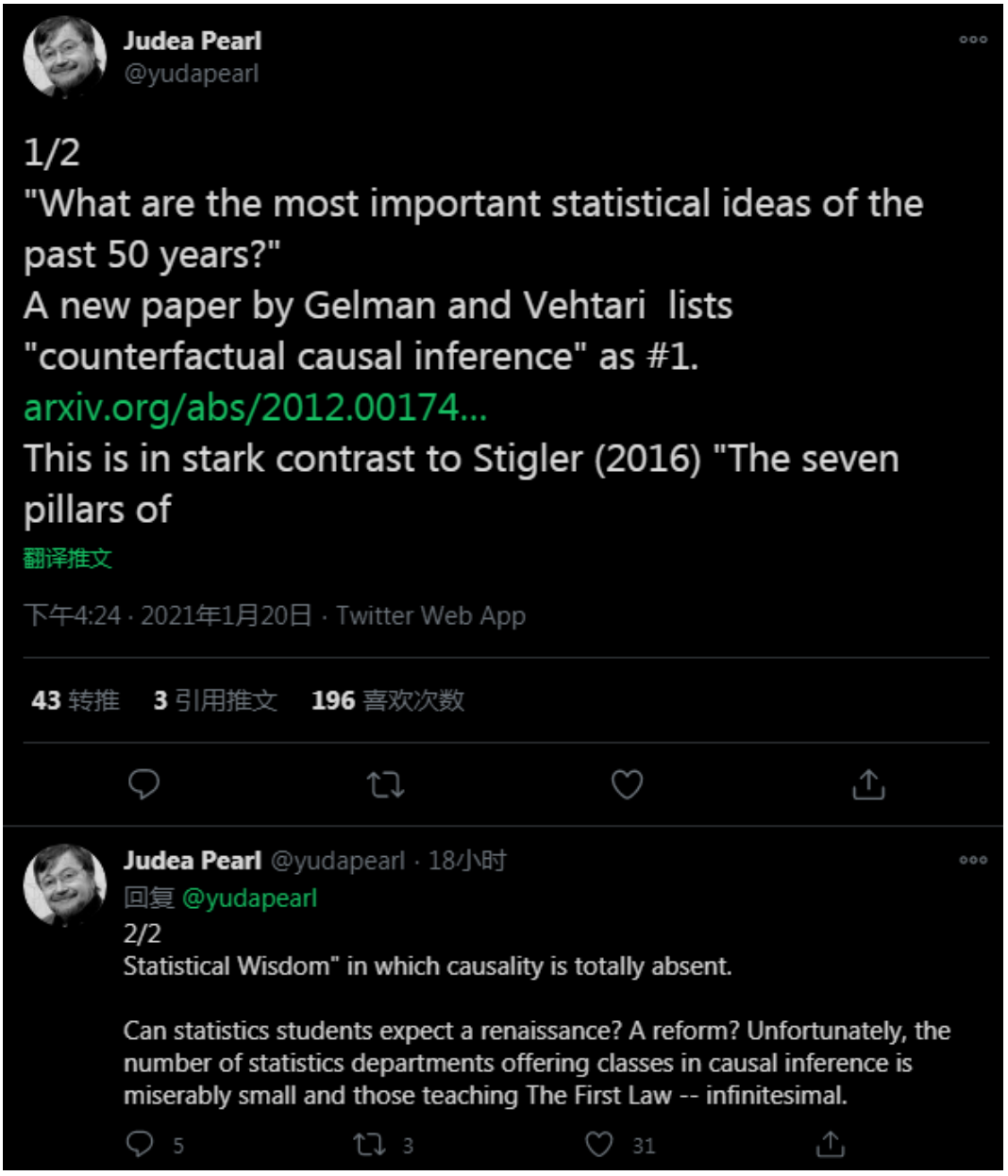
近日，圖靈獎得主、“貝葉斯網路之父”Judea Pearl在Twitter上分享了一篇新論文“What are the most important statistical ideas of the past 50 years?”（過去50年中最重要的統計思想是什麼？）

這篇論文由哥倫比亞大學統計學教授Andrew Gelman和阿爾託大學計算機科學系副教授Aki Vehtari所著，他們根據自己的研究和文獻閱讀經驗總結出了**過去半個世紀以來最重要的8個統計思想**，並表示：“它們是獨立的概念，涵蓋了統計方面不同的發展。這些思想都在1970年前的理論統計文獻和各個應用領域的實踐中就已經出現。但是在過去的五十年中，它們各自已經發展到足以成為新事物的程度。”

他們認為，過去半個世紀中最重要的統計思想是：**反事實因果推理，基於bootstrapping（自助抽樣法）和基於模擬的推理，超引數化模型和正則化，多層模型，泛型計算演算法（generic computation algorithms），自適應決策分析，魯棒推理和探索性資料分析**（未按時間順序，排序不分先後）。

在這篇論文中，**他們將討論這些思想的共同特徵、它們與現代計算和大資料的關係以及在未來幾十年中如何發展**。“本文的目的是引起有關統計和資料科學研究更大主題的思考和討論。”

值得一提的是，Judea Pearl在推文中表示，“對作者將因果推理列入其中感到欣慰，這與Stigler在《統計學七支柱》中的總結截然不同，後者完全沒有提到因果推理。”另外，他也對大學統計專業很少安排因果推理課程感到擔憂，“統計學可以期待復興或改革嗎？不幸的是，統計系中提供因果推理課程的非常少，更不用提教'The First Law'的，簡直是無窮少。”



论文：What are the most important statistical ideas of the past 50 years?

論文地址：https://arxiv.org/pdf/2012.00174.pdf

作者簡介：

Andrew Gelman，美國統計學家，哥倫比亞大學統計學和政治學教授。他1986年獲得麻省理工學院數學和物理學博士學位。隨後，他獲得了博士學位。在哈佛大學統計學榮譽退休教授Donald Rubin的指導下，於1990年從哈佛大學獲得統計學博士學位。他是美國統計協會與數理統計學會的院士，曾三度獲得美國統計協會頒發的“傑出統計應用獎”，谷歌學術顯示，他的論文總引用量超過12萬，h-index為110。

Aki Vehtari，阿爾託大學計算機科學系副教授，主要研究領域為貝葉斯機率理論和方法、貝葉斯工作流、機率程式設計、推理方法（例如Laplace，EP，VB，MC）、推理和模型診斷、模型評估和選擇、高斯過程以及分層模型。谷歌學術顯示，他的論文總引用量近4萬。他和Andrew Gelman都是《貝葉斯資料分析》的作者，這本書因在資料分析、研究解決難題方面的可讀性、實用性而廣受讀者好評，被認為是貝葉斯方法領域的優秀之作。

以下是全文編譯：

1、過去50年最重要的統計思想

1.1 反事實因果推理

在這裡，我們首先要介紹在統計學、計量經濟學、心理測量、流行病學和計算機科學領域出現的一些重要思想，它們都圍繞著因果推理面臨的挑戰展開，並且都在某種程度上彌平了「對觀測推理的因果解釋」和「認識到關聯關係並不意味著因果關係」這兩方面的差距。

核心的思想在於，在某些假設情況下，我們可以識別出因果關係，而且我們可以嚴謹地宣告這些假設，並且透過設計和分析以各種方式解決它們。

到目前為止，關於如何將因果模型應用於真實資料的具體問題上的爭論仍在繼續。然而，在過去的五十年中，這一領域的工作進展使因果推理所需要的這些假設變得精確得多，從而反過來又促進了解決這些問題的統計方法的相關工作。

研究人員針對各個領域研發出了各種各樣的因果推理方法：在計量經濟學領域中，人們主要關注對線性模型的因果估計的可解釋性；在流行病學領域中，人們主要關注基於觀測資料的推理；心理學家已經意識到互動和各種處理效應的重要性；在統計學領域中，出現了一系列有關匹配和其它調整並衡量實驗組和對照組之間差別的方法；在計算機科學領域中，湧現出有關多維因果歸隱模型的研究工作。

在上述所有工作中，有一條研究主線，**即從反事實或可能得到的結果的層面上對因果問題進行建模，這相較於之前沒有明確區分描述性推理和因果推理的標準方法是一個巨大的飛躍。**

在這個研究方向上，具有里程碑意義的工作包括 Neyman (1923)，Welch (1937)，Rubin (1974)，Haavelmo (1973) 等人的研究成果，更加詳細的研究背景請參閱 Heckman 和 Pinto 於 2015 年發表的論文「Causal analysis after Haavelmo」。

反事實因果推理的思想和方法在統計學以及相關的應用研究和策略分析領域都有深遠影響。

1.2 bootstrap與基於模擬的推理

在過去的50年中，用計算取代數學分析是統計學的一大發展趨勢。這一變化甚至在「大資料」分析出現之前就開始了。

bootstrap是最純粹的基於計算定義的統計方法之一，它定義了一些估計量，並將其應用於一組隨機重取樣資料集。**其思想是將估計值視為資料的一個近似的充分統計量，並將自助分佈視為對資料的取樣分佈的近似。**在概念層面上，人們推崇將預測和重新抽樣作為基本原則，可以推匯出諸如偏差校正和收縮等統計學操作。

歷史上，這一方向誕生了「刀切法」和「交叉驗證」等方法。此外，由於bootstrap思想的通用性及其簡單的計算實現方式，bootstrap立刻被廣泛用於各種傳統的解析近似方法效果不佳應用，從而產生了巨大的影響。**時至今日，充足的計算資源也起到了幫助作用，使得對許多重取樣得到的資料集進行反覆的推理變得十分容易。**

計算資源的增加也使得其它重取樣和基於模擬的方法流行了起來。在置換檢驗中，我們透過隨機打亂排列真實值（target）來打破預測值和真實值之間的依賴關係，從而生成重取樣資料集。引數化的bootstrap、先驗和後驗預測檢查、基於模擬的校正都是根據模型建立了複製資料集，而不是直接從資料中重取樣。在分析複雜模型和演算法時，根據已知的資料生成機制取樣的做法往往被用於建立模擬實驗，用於補充或替代數學理論。

1.3 過引數化模型和正則化

自 20 世紀 70 年代以來，統計學受個方面的影響，發生了一個重大的變化，即用一些正則化過程得到穩定的估計和良好的預測結果，從而擬合具有大量引數（有時引數比資料點更多）的模型。**該思想旨在在避免過擬合問題的同時，獲得一種非引數化的或高度引數化的方法。**我們可以透過針對引數或預測曲線的懲罰函式來實現正則化。

早期的高度引數化的模型包括「馬爾科夫隨機場」、「樣條函式」、「高斯過程」，隨後又出現了「分類和迴歸決策樹」、「神經網路」、「小波收縮」、「Lasso 和 Horseshoe 等最小二乘的替代方法」、「支援向量機及相關理論」。

上述所有模型都會隨著樣本規模的增加而擴大，其引數往往也不能被直接解釋，它們是一個更大的預測系統的一部分。在貝葉斯方法中，我們可以首先在函式空間中考慮先驗，然後間接推匯出相應的模型引數的先驗。

在人們能夠容易地獲得充足的計算資源之前，這些模型的使用還十分有限。此後，影象識別、深度神經網路領域中的過引數化模型持續發展。Hastie、Tibshirani 以及 Wainwright 於 2015 年將許多該領域的工作定義為對稀疏結構的估計。

但是在本文作者看來，**正則化技術更為通用，這是因為它使稠密的模型能夠適應資料支援的程度。**在統計學領域以外， 這方面也產出了許多成果，例如：非負矩陣分解、非線性降維、生成對抗網路、自編碼器。它們都是可以尋找結構和分解結果的無監督學習方法。

隨著統計方法得到了發展，並被應用於更大的資料集上，研究者們還研發了一些調優、自適應，以及組合來自多個擬合結果的推理（包括 stacking 整合、貝葉斯模型平均、boosting 整合、梯度提升、隨機森林）的方法。

1.4 多層模型

多層模型的引數因組而異，它使模型可以適應於聚類抽樣、縱向研究、時間序列橫斷面資料、元分析以及其它結構化的環境。在迴歸問題中，一個多層模型可以被看做特

定引數化的協方差結構，或者是一個引數數量隨資料比例增加的機率分佈。

多層模型可以被看做一種貝葉斯模型，它們包含未知潛在特徵或變化引數的機率分佈。反過來，貝葉斯模型也有一種多層結構，包含給定引數的資料和超引數的引數的分佈。

對區域性和一般資訊進行池化（pooling）的思想是根據帶有噪聲的資料進行預測的固有數學原理。這一思想可以追溯到拉普拉斯和高斯，高爾頓也隱式地表達了這種思想。

部分池化的思想已經被應用於一些特定應用領域（例如：動物育種）。部分池化與統計估計問題中的多重性的一般關係由於 James 和 Stein 等人的工作而得到了理論上的重要進展。最終，這啟發了心理學、藥理學、抽樣調查等領域的研究。Lindley 和 Smith 於 1972 年發表的論文，以及 Lindley 和 Novick 於 1981 年發表的論文提供了一種基於估計多變數正態分佈的超引數的數學結構，而 Efron 和 Morris 等人則給出了相應的決策理論方面的解釋，接著這些思想被融入了迴歸建模並被應用於廣泛的使用結構化資料的問題。

從另一個方向來看，Donoho 等人於 1995 年給出了多元引數收縮的資訊理論解釋。**我們更傾向於將多層模型看做將不同的資訊源進行組合的框架，而不是一個特定的統計模型或計算過程。**因此，每當我們想要根據資料的子集進行推理（小面積估計）或將資料泛化到新問題（元分析）上的時候，就可以使用這種模型。類似地，**貝葉斯推理的可貴之處在於，它不僅僅是一種將先驗資訊和資料組合起來的方法，也是一種解釋推理和決策的不確定性的方法。**

1.5 泛型計算方法

前文中討論過的建模方面的研究進展高度依賴於現代計算科學，這不僅僅指的是更大的記憶體、更快的 CPU、高效的矩陣計算、對使用者友好的語言，以及其它計算科學方面的創新。用於高效計算的統計算法方面的進展也是一個關鍵的因素。

在過去的 50 年中，在統計問題的結構方面出現了許多具有創新性的統計算法。EM 演算法、Gibbs 取樣、粒子濾波、變分推斷、期望傳播以不同的方式利用了統計模型的條件獨立結構。

而 Metropolis 演算法、混合或 Hamiltonian 蒙特卡洛演算法則並沒有直接受到統計問題的啟發，它們最初被提出用於計算物理學中的高維機率分佈，但是它們已經適應了統計計算，這與在更早的時候被用於計算最小二乘以及最大似然估計的最佳化演算法相同。

當似然的解析形式很難求解或計算開銷非常大時，被稱為近似貝葉斯計算的方法（透過生成式模型模擬、而不是對似然函式進行估計得到後驗推理）是十分有效的。

縱觀統計學的歷史，資料分析的發展、機率建模和計算科學是共同發展的。新的模型會激發具有創新性的計算演算法，而新的計算技術又為更加複雜的模型和新的推理思想開啟了大門（例如，高維正則化、多層建模、自助抽樣法）。通用的自動推理演算法使我們可以將模型的研發解耦開來，這樣一來變更模型並不需要對演算法實現進行改變。

1.6 自適應決策分析

自 20 世紀 40 年代至 20 世紀 60 年代，決策理論往往被認為是統計學的基石，代表性的工作包括：效用最大化、錯誤率控制、以及經驗貝葉斯分析。

近年來，沿著上述工作的方向，研究人員在貝葉斯決策理論、錯誤發現率分析等領域也取得了一系列成果。決策理論還受到了有關人類決策中的啟發與偏見的心理學研究的影響。

決策也是統計學的應用領域之一。**在統計決策分析領域的領域中，重要的研究成果包括：貝葉斯最佳化、強化學習**，這與工業中的 A/B 測試的實驗設計的復興以及許多工程應用中的線上學習有關。

計算科學的最新進展使我們可以將高斯過程和神經網路這些高度引數化的模型用作自適應決策分析中的函式的先驗，還可以在模擬環境中進行大規模的強化學習，例如：創造能夠控制機器人、生成文字、以及參與圍棋等遊戲。

1.7 魯棒的推理

魯棒性思想是現代統計學的核心，它指的是：即使在假設錯誤的前提條件下，我們也可以使用模型。實際上，開發出能夠在違背上述假設的真實場景下良好執行的模型對於統計理論來說是十分重要的。

Tukey 曾於 1960 年在論文「A survey of sampling from contaminated distributions」中對該領域的工作進行了綜述，Stigler 也於 2010 年在論文「The changing history of robustness」中進行了回顧。

受到 Huber 等人工作的影響，研究者們開發出了一系列在現實生活中（尤其是經濟學領域，人們對統計模型的缺陷有深刻的認識）具有一定影響力的魯棒方法。在經濟學理論中，存在「as if」分析和簡化模型的概念，因此計量經濟學家會對在一系列假設下還能執行良好的統計程式十分感興趣。例如，經濟學和其它社會科學領域的應用研究人員廣泛使用魯棒標準誤差以及部分識別。

一般來說，正如在 Bernardo 和 Smith 於 1994 年所提出的「M-開放世界」（在這個世界中，資料生成過程不屬於擬合的機率模型）下評估統計過程的想法一樣，統計研究中的魯棒性的主要影響並不在於對特定方法的發展。Greenland 認為，研究者需要顯式地解釋傳統統計模型中沒有考慮的誤差來源。**對魯棒性的關注與高度引數化的模型相關，這是現代統計學的特點，對模型評估有更普遍的影響。**

1.8 探索性資料分析

上文討論的統計思想都涉及密集的理論和計算的結合。從另一個完全不同的方向來看，研究人員們進行了一種具有影響力的「迴歸到本質」的探索，**他們跳出機率模型，重點關注資料的圖形視覺化。**

Tukey 和 Tufte 等人在他們的著作中曾對統計圖的優點進行了討論，而許多這樣的思想透過他們在資料分析環境 S（目前在統計學及其應用領域佔據主導地位的 R 語言的前身）中的實現開展了統計實踐。

在 Tukey 之後，探索性資料分析的擁躉重點說明了漸進理論的侷限性以及開放式探索和通訊的好處，並且闡明瞭超越統計理論的對統計科學的更一般的觀點。**這與更加關注發現而非檢驗固定假設的統計建模觀點不謀而合。**

同時，它不僅在特定的圖形化方法的發展中十分具有影響力，也從科學的資料中學習，將統計學從定理證明轉向更開放、更健康的角度。舉例而言，在醫學統計學領域中，Bland 和 Altman 於 1986 年發表的一篇高被引論文推薦人們將圖形化方法用於資料對比，從而替換關聯性和迴歸分析。

此外，研究人員試圖形式化定義探索性資料分析：Gelman 將資料展示與貝葉斯預測檢查的視覺化相結合，Wilkinson 形式化定義了統計圖中固有的對比和資料結構，而 Wickham 透過這種方式得以實現了一個極具影響力的 R 語言程式包，從而在許多領域中改變了統計學實踐。

計算的進步使從業者們能夠快速構建大型的複雜模型，其中在理解資料、擬合的模型、預測結果之間的關係時，統計圖是十分有用的。「探索性模型分析」有時被用來獲取資料分析過程的實驗特性。研究人員們也一直進行著將視覺化囊括在模型構建和資料分析過程中的研究工作。

2、相同點與不同點

2.1 思想能產生方法與工作流程

我們之所以認為上面列出的思想重要，是因為它們不僅解決了現有問題，還建立了新的統計思維方式和資料分析方式。換句話說，上述的每一種思想都是一部法典，其方法不僅侷限於統計學，而更像是一種“研究品味”或“哲學思想”：

- 反事實機制將因果推理置於統計或預測的框架中，其中，因果估量（causal estimands）可以根據統計模型中未觀察到的資料精確定義和表達，並與調查抽樣和缺失資料推算的思想聯絡起來。
- Boostrap 打開了隱式非引數建模（implicit nonparametric modeling）的大門。
- 過引數化的模型和正則化基於從資料中估計模型引數的能力，將限制模型大小的現有做法形式化和泛化，這與交叉驗證和資訊標準有關。

- 多層模型將從資料估計先驗分佈的“經驗貝葉斯”技術形式化，使這種方法在類別更廣泛的問題中使用時具備更高的計算與推理穩定性。
- 泛型計算演算法使實踐者能夠快速擬合用於因果推理、多層次分析、強化學習和其他許多領域的高階模型，使核心思想在統計學和機器學習中產生更廣泛的影響。
- 自適應決策分析將最佳控制的工程問題與統計學習領域聯絡在一起，遠遠超出了經典的實驗設計。
- 魯棒推理將對推理穩定性的直覺形式化，在表達這些問題時可以對不同程式進行正式評估和建模，以處理對異常值和模型錯誤說明的潛在擔憂。此外，魯棒推理的思想也為非引數估計提供了資訊。
- 探索性資料分析使圖形技術和發現成為統計實踐的主流，因為這些工具正好可以用於更好地理解和診斷正在與資料進行擬合的機率模型的新型複雜類別。

2.2. 計算上的進步

元演算法（利用現有模型和推理步驟的工作流）在統計學中被廣泛使用，比如最小二乘法，矩估計（the method of moments），最大似然，等等。

在過去 50 年裡所開發的許多機器學習元演算法都有一個特徵，就是它們會以某種方式拆分資料或模型。學習元演算法（Learning Meta-Algorithms）與分治計算方法相關，最著名的是變分貝葉斯和期望傳播演算法。

元演算法和迭代計算在統計學中之所以重要，主要是有兩個原因：1）除了最初開發的元演算法示例以外，透過多個來源整合資訊，或透過整合弱分類器（weak learner）來建立強分類器的通用想法可以得到廣泛應用；2）自適應演算法在線上學習中發揮了很好的作用，最終被認為代表了現代統計觀點：資料和計算分開，資訊交換和計算架構是元模型或推理過程的一部分。

新方法使用新技術工具並不稀奇：隨著計算速度越快、計算範圍越廣，統計學家不再侷限於具備解析方案的簡單模型與簡單的封閉式演算法（如最小二乘法）。我們可以簡要說一下上述思想是如何利用現代計算：

- 一些思想（boostrapping，超引數化模型和機器學習元分析）直接利用了計算速度，這在計算機出現之前難以想象。例如，直到引入高效的GPU卡和雲計算之後，神經網路才更加流行起來。
- 除了計算能力以外，計算資源的分散也很重要：臺式計算機能讓統計學家和計算機科學家嘗試新方法，然後由從業人員使用這些新方法。
- 探索性資料分析最初是從紙筆圖形開始，但隨著計算機圖形學的發展，探索性資料分析已經歷徹底改變。

- 過去，貝葉斯推理僅限於可以透過分析解決的簡單模型。隨著計算能力的提高，變分和馬爾可夫鏈模擬方法使得模型構建和推理演算法開發的分離成為可能，機率程式設計也因此允許不同領域的專家能夠專注於模型構建並自動完成推理。這導致了貝葉斯方法在1990年開始在許多應用領域變得普及。
- 自適應決策分析，貝葉斯最佳化和線上學習應用於計算和資料密集型問題，例如最佳化大型機器學習和神經網路模型，實時影象處理和自然語言處理。
- 魯棒的統計學不一定需要大量計算，但它的使用在一定程度上由計算驅動，與封閉式估計（如最小二乘法）有所區別。Andrews等人曾使用大量計算進行了一項模擬研究，促進了對魯棒方法的開發和理解。
- 減少多元推理的合理性不僅可以透過統計效率來證明，還可以從計算層面證明：激發了一種新的漸近理論。
- 反事實因果推理的關鍵思想與理論相關，而不是計算相關。但是，近年來，因果推理在使用計算密集的非引數方法後已有了發展，促進了統計學、經濟學和機器學習中因果和預測模型的統一。

2.3 大數據

除了為統計分析開拓發展空間以外，現代計算還啟發了新統計方法的應用和開發，從而產生了大資料，例子有：基因陣列，流影象和文字資料，以及線上控制問題，如自動駕駛汽車。**事實上，“資料科學”流行的原因之一就是因為，在此類問題中，資料處理和高效計算是與用於擬合數據的統計方法一樣重要的。**

這與 Hal Stern 的觀點相關：**統計分析最重要的方面不是對資料進行的操作，而是你所使用的資料是什麼。**與先前的方法相比，本文討論的所有思想都有一個共同特徵，即有助於使用更多的資料：

- 反事實框架允許使用用於對受控實驗建模的相同結構從觀測資料中進行因果推斷。
- Bootstrapping 可用於糾正偏差，與在分析計算無法進行的複雜調查、實驗設計和其他資料結構上進行方差估計。
- 正則化允許使用者在模型中加入更多預測變數，而不必擔心過度擬合。
- 多層模型使用部分彙集來合併來源不同的資訊，從而更廣泛應用元分析的原理。
- 泛型計算演算法允許使用者擬合更大的模型，這對將可用資料連線到重要的基本問題來說可能是有必要的。
- 自適應決策分析利用在數值分析中開發的隨機最佳化方法。
- 魯棒推理可以更常規地使用具有異常值、相關性和其他可能阻礙常規統計建模的資料。

- 探索性資料分析為複雜資料集的視覺化打開了大門，並推動了整潔資料分析（tidy data analysis）的發展，以及統計分析、計算和通訊的整合。

在過去的50 年裡，統計程式設計環境也有了很大的發展，最著名的是S語言、R語言，還有以 BUGS 開頭命名的通用推理引擎及其後繼者。近日，數值分析、自動推理和統計計算的思想開始以可複製的研究環境（如Jupyter notebook）和機率程式設計環境（如Stan，Tensorflow和Pyro）的形式混合在一起。因此，我們至少可以預計推理和計算方法的部分統一，例如使用自動微分進行最佳化、取樣和靈敏度分析。

2.4 這些思想的關聯與互動

Stigler 在 2016 年提出，**一些明顯不同的統計領域背後存在某些共同主題的相關性。**這一互相聯絡的思想也可以用於最近的發展。例如，魯棒統計學（側重於偏離特定模型假設）和探索性資料分析（傳統上被認為對模型根本不感興趣）之間有什麼聯絡？

探索性方法（如殘差圖和 hanging rootograms ）可以從特定的模型分類（分別是累計迴歸和泊松分佈）中獲得，但是，它們的價值在很大程度上是在於其可解釋性，即無需參考啟發它們的模型。

同樣，你可以單獨將一種方法（如最小二乘法）看作對資料的運算，然後研究表現好的資料生成過程的類別，再使用這種理論分析的結果來提出更魯棒的程式，能夠拓展無論是基於故障點（breakdown point），極小化極大風險或其他方式定義的適用範圍。相反，純粹的計算方法（例如蒙特卡洛積分估算）可以被有效解釋為統計推理問題的解決方案。

另一個聯絡是，因果推理的潛在結果框架對人群中的每個單元都有不同的處理效應，因此自然而然就採用了一種元分析方法將效應多樣化，並使用在實驗或觀察性研究分析中使用多層次迴歸進行建模。

回過頭來看，研究 bootstrap 可以為我們提供一種新觀點：將經驗貝葉斯（多層次）推理看作非透視方法。在該方法中，正態分佈或其他引數模型用於部分彙集，但最終估計值不侷限於任何引數形式。對小波（wavelets）和其他豐富引數化模型進行正則化的研究與在魯棒背景下開發的穩定推理程式之間存在意想不到的聯絡。

其他方法論的聯絡更為明顯。正則化的過引數化模型使用機器學習元演算法進行了最佳化，反過來又可以得出對 contamination 具有魯棒性的推論。這些連線可以用其他方式表示，魯棒迴歸模型對應混合分佈，混合分佈可以視為多層次模型，還可以使用貝葉斯推理進行擬合。深度學習模型與一種多層次邏輯迴歸相關，也與復現核心的 Hilbert 空間（在樣條中使用，支援向量機）相關。

高度引數化的機器學習方法可以構建為貝葉斯分層模型，其中將懲罰函式正則化與超先驗相一致，無監督學習模型也可以被構建為具有未知組員的混合模型。在許多情況下，是否使用貝葉斯生成框架是取決於計算，這也是雙向進行：貝葉斯計算方法可以

幫助掌握推理和預測中的不確定性，高效最佳化演算法也可以用於近似基於模型的推理 。

許多被廣泛討論的思想都涉及到豐富的引數化，並伴隨一些用於正則化的統計或計算工具。因此，它們可以被認為是經篩選思想的更廣泛實現：隨著可用資料的增加，模型會變得更大。

2.5 理論促進應用，反之亦然

可以說所有這些方法的共同特徵是易記的名稱和良好的傳播。但是作者懷疑這些方法的名稱僅在回顧時會引起注意。諸如“反事實”、“載入程式”、“堆疊”和“增強”之類的術語聽起來很專業，而不是令人印象深刻，作者認為**是方法的價值使這些名字變得響亮。**

創新的想法經常會遇到阻力，這也是本文中討論的這些有影響力的想法的命運。**如果一個新思想起源於一個應用領域，那麼要說服理論家相信它的價值可能會遇到很大挑戰。相反，批評新方法在理論上是有用的，但在實踐中沒有用，倒是很容易。**

我們應該澄清，所謂“反對”不一定意味著積極反對。**與其他一些學術領域相比，統計資料不是很政治化：**學術界、政府和行業內部對統計領域的發展很寬容，甚至邊緣思想也被允許發展。此處討論的許多方法（例如載入程式，lasso和多層模型）在統計和各種應用領域中都立即流行起來，但即使是這些思想也面臨著阻力，即局外人需要確信其應用的必要性。

理論統計學是應用統計學的理論，這在一定程度上得益於諸如Cox的“Planning of Experiments”，Box and Tiao的“Bayesian Inference in Statistical Analysis”，Cox and Hinkley的“Theoretical Statistics”，Box，Hunter和Hunter的“Statistics for Experimenters”等有影響力的著作，幫助我們跨越了理論和應用之間的鴻溝。

不同於純數學，不存在純粹的統計。沒錯，一些統計思想是深刻而優美的，並且與數學一樣，這些思想也具有基本的聯絡。例如，迴歸和均值之間的聯絡，最小二乘和部分池化之間的聯絡，但它們仍與特定主題相關。就像摘下的蘋果一樣，脫離其營養來源後，理論統計研究趨於枯竭。數學也是如此，但是純數學中的思想似乎可以存在更長的時間，並且能以孤立的研究存在，而統計學思想則無法如此。

應用統計理論帶來的好處是顯而易見的。**人們可以將理論視為計算的捷徑。**我們總是需要這樣的捷徑：建模的需求不可避免地隨著計算能力的增長而增加，因此我們需要分析壓縮和逼近的價值。此外，理論可以幫助我們理解統計方法的工作原理，而數學邏輯可以啟發新的模型和資料分析方法。

2.6 和統計領域其他進展的關聯

特定的統計模型與這些重要思想是什麼聯絡？在這裡，作者考慮的是有影響力的工作，例如風險迴歸、廣義線性模型、空間自迴歸、結構方程模型、潛在分類、高斯過

程和深度學習。如上所述，在過去的半個世紀中，統計推斷和計算領域出現了許多重要的發展，這些發展都受到了上面討論的新模型和推斷思想的啟發和推動。**模型、方法、應用程式和計算都結合在一起**。

討論不同概念發展之間的聯絡，並不意味著關於適當使用和解釋統計方法的爭論仍然存在。例如，錯誤發現率（false discovery rate）與多層模型之間存在雙重性，但是基於這些不同原理的過程可以給出不同的結果。通常使用貝葉斯方法來擬合多層模型，並且在後驗分佈中，沒有任何東西會一直收斂到零。

相反，錯誤發現率方法通常使用p值閾值，目的是識別少量統計上顯著的非零結果。再例如，在因果推理中，人們越來越關注密集引數化的機器學習預測，然後進行後分層（poststratification）以獲得特定的因果估計，但是在更開放的環境中，需要發現非零因果關係。同樣，根據目標是密集預測還是稀疏預測，使用了不同的方法。

最後，我們可以將統計方法的研究與科學和工程學中統計應用的趨勢聯絡起來。在這裡，**作者提到了生物學、心理學、經濟學和其他科學領域的復現危機或可復現性革命，這些領域的變異範圍足夠大，需要根據統計證據得出結論。**

在可復現性革命中，具有里程碑意義的論文包括：

Meehl發表的“Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology”，概述了在原假設重要性檢驗的標準用法中提出科學主張的哲學缺陷。

Ioannidis發表的“Why most published research findings are false”，其認為，醫學上大多數已發表的研究都在使得結論不受其統計資料的支援。

Simmons，Nelson和Simonsohn發表的“False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant ”，解釋了“研究人員的自由度”如何使研究人員即使從純噪聲資料中也能常規獲得統計意義。

一些補救措施是程式性的，例如Amrhein，Greenland和McShane發表的“Scientists rise up against statistical significance”。

但也有人建議可以使用多層模型解決不可復現研究的某些問題，將估計值部分歸零以更好地反映研究中的效應總量，例如van Zwet，Schwab和Senn發表的“The statistical properties of RCTs and a proposal for shrinkage”。

可再現性和穩定性問題也直接涉及到載入程式和可靠的統計資料，參見Yu. B.發表的“Stability.”。

3、未來幾十年的重要統計思想會是什麼？

3.1 回顧

在考慮自1970年以來最重要的發展時，回顧一下1920–1970年的重要統計思想（包括質量控制、潛在變數建模、抽樣理論、實驗設計、經典和貝葉斯決策分析、置信區間和假設檢驗、最大似然、方差分析和客觀貝葉斯推理）也很有意義。當然還有1870年至1920年（機率分佈分類、均值迴歸、資料現象學建模），以及Stigler在《The History of Statistics》中提到的更早年代的統計思想。

在本文中，作者試圖提供一個廣泛的視角，以反映不同的觀點。但是其他人可能對過去五十年來最重要的統計思想有自己的看法。確實，問這個問題主要是引起人們對統計學觀念的重要性的討論。在本文中，作者避免了使用引文計數或其他數值方法對論文進行排名，但是他們隱含地以類似page-rank的方式來衡量影響力，因為他們試圖將注意力集中在那些影響了統計實踐的方法發展的思想上。

3.2 展望

接下來會發生什麼？**作者同意卡爾·波普爾（Karl Popper）的觀點，即人們無法預見所有未來的科學發展，但是我們可能對當前的趨勢將如何持續有比較可靠的見解。**

最安全的選擇是，在現有方法組合上持續取得進展：對潛在輸出的豐富模型進行因果推理，並使用正則化估計；結構化資料的複雜模型，例如隨時間演變的網路，對多層模型的可靠推斷；對超引數化模型的探索性資料分析；用於不同計算問題的子集（subsetting）和機器學習元演算法等等。此外，作者期望在結構化資料的實驗設計和取樣方面取得進展。

另一個成熟的發展領域是模型理解，有時也稱為可解釋機器學習。這裡的矛盾之處在於，理解複雜模型的最佳方法通常是使用簡單模型對其進行近似。但問題是，在這過程中是什麼在進行交流？一種可能有用的方法是計算對資料和模型引數擾動的推斷敏感性，將魯棒性和正則化的思想與基於梯度的計算方法相結合，該方法在許多不同的統計算法中使用。

最後，鑑於幾乎所有新的統計和資料科學思想在計算上都是昂貴的，因此，作者設想了對推論方法驗證的未來研究，將諸如軟體工程中的單元測試之類的思想應用到從噪聲資料中學習的問題中。**隨著統計方法變得越來越先進，理解資料、模型和實體理論之間的聯絡將變得越來越重要。**

※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※

我是「數據分析那些事」。常年分享數據分析乾貨，不定期分享好用的職場技能工具。各位也可以關注我的Facebook，按讚我的臉書並私訊「10」，送你十週入門數據分析電子書唷！期待你與我互動起來~

文章推薦

◆SQL零基礎入門必知必會！

◆如何系統地學習資料探勘？