

# 統計是數據科學<sup>1</sup>

趙民德<sup>2</sup>

## 前言

這篇文章應該用較長的標題：統計學的發展，是根據數據的形式的改變而發展的。統計是不是處理數據的科學，可以爭論。但是，較短的標題，比較有力。

科學的發展有兩個源頭：好奇和需要。別的科学如此，統計科學亦然。好奇，是有了相當程度之後的事。因為一般的好奇，深度不夠。而深層的好奇，則來自切實的了解和精準的命題。純為了解決實際問題而導出的學問，如果不繼續深究，最多只是科技。但好奇也容易流入孤軍深入，後援不繼的困境。

統計科學既是入世，也是出世。它因為數據的發生而產生，因我們對於某些觀念——如誤差——的意義而深入 (Stigler, 1986)。多半人看到的統計學，只是一些彙總後的數字摘要，再進一步，則是整本整冊的報表；多半研究者所看到的統計學，又充滿了數學符號和公式。因此一個統計專業的畢業生，不論是學士、碩士還是博士，無可避免地在一般人所了解的統計和他所受到訓練的統計間，有一點無所適從。

統計的本質是它所分析的數據：我們都是先看到數據，再想到如何去分析。古人如此，現代人也應該如此。只是現在的統計學裡，已有太多的前人已發展的模型。這就像已有一般解的定論相仿：破當頭炮，就得用屏風馬。現在你學過的大多是現成已有的招式：ANOVA，迴歸，對數線性模型...

招式是多年經驗累積下來的。但是統計的精神，則在數據結構的本身和所想要得到的解答。在科學和實務上，你不一定遇到教本上所描繪的數據格式的條件和數量。因此統計方法要一再針對新型的數據和問題來加以調適。而在某一種新型數據開始大量出現並亟力要求統計界加以詮註的時候，新的統計方法和理論就會發展出來。

從統計發展幾百年的歷史來看，這大概是一個基本法則。數據之外，或者尚有其它的要求——例如管理者的希望和意識型態的作祟——但是統計科學，作為一種科學，無可避免地將一再回歸它的本質：數據。

本文從這樣的角度，來探討統計科學的發展。

---

<sup>1</sup>此稿的不同形式，講過幾次。最早也許是在高雄大學，也在人民大學講過，是我預備的《統計十講》之一，後來並刊在《中國統計》的某一期。

<sup>2</sup>中研院統計所，已退休。

## 因政令的需求而有統計

古早的統計我們不談<sup>3</sup>，但世界上範圍最廣的統計專業學社：總部設在荷蘭的國際統計學院（International Statistics Institute, ISI）就有超過一百年（1885 - ）的歷史。古早的統計來自施政（如抽稅和征兵）的需求，時至今日，政府統計仍然在各國的統計專業團體裡佔有舉足輕重的地位。因為它常川而定期並專業地蒐集資訊，是任何有責任心的管理單位都不能忽視的工作。

政府的統計工作，是以蒐集、整理、發布為骨幹。它不止是在看全國，也要看年與年與間的變化。所以橫截面（cross sectional）數據和時間序列（time series）數據，便成為主流。但因為這種類型的數據在型態上變化較小，且對於它們的處理方法，不論是做得好還是不好，都至少有五六十年以上歷史的延續發展，且都已在各國的政府統計中到位，在一般的統計研究中，在方法論方面，較難有令人興奮的結果。

政府統計的主要意義仍然在於其發布的本質以及對於官方意見的立場表示。即便做得不盡理想，但官方的數字仍有其一定的影響力。因為它一方面代表該國的狀況，二來也可能是世上唯一的相關資料<sup>4</sup>。但從另一方面來說，政府統計都有一定的時空背景。例如 1929<sup>5</sup> 年的資料，不論做得多麼好，不論它是 CPI, GDP, 個人平均所得還是產出投入表，大概都不會得到太多的關注。因此，長期下來，我們大概只能看到最後的數字，而對於其所用的方法，不論是蒐集、整理和分析，都了解不深。

這是所有實務系統都有的問題。政府統計因為在地的意義、官方的色彩及相伴而來的不求有功但求無過的文官系統的運作，問題自然更為嚴重。一般的統計學家想要真的了解政府統計並不容易，而不同政府間的政府統計，也未見得會用相同的語言。雖然，聯合國統計處的存在，讓定義不盡相同的問題少了一點，但未必根絕。

新的事物當然仍然在開發中。當政府部門在大量使用電腦的時候<sup>6</sup>，當社會對隱私權日益重視的時候，當民衆對環境和生活品質要求日高的時候，政府統計的方法和態度，當然也會跟著調整。但是大體上，政府統計和經濟這個領域密切掛鉤，是大約不會變的。因為政府的目的是讓人民活得好，而經濟學用較全面的角度來關注這樣的問題。當然，環境、衛生、生活品質、社會安全這些概念，也會早晚一一會反應在政府統計裡。

政府統計的骨幹，是一套有效率且精良的資料蒐集和呈報的系統。而這套學問是很

---

<sup>3</sup>例如我們可以上溯到《管子》的《牧民篇》，這也許是中國最早談到統計的文獻。

<sup>4</sup>意思是說，其它的單位，很難獨立地再做同樣的工作，因此官方資料一般都缺乏獨立的佐證。

<sup>5</sup>這是中國開始有統計制度的一年。

<sup>6</sup>最早使用電腦的政府單位，是美國普查局。

難在學校裡學到的。因為沒有一個學校會有那樣的規模。但是，至少抽樣學的基本理論仍然一再地被使用著。

## 抽樣

抽樣是以偏論全的技術。在 1934 年以前<sup>7</sup>，抽樣是以立意抽樣為主：抽樣者主觀地去選擇一組他認為有代表性的樣本。但在 1934 年之後，以機率論為基礎的隨機抽樣逐漸地取得了抽樣的主流地位。抽樣的目的極為明顯地和數據相關：有效地取得數據。隨機抽樣之所以能夠大行其道，主要是因為它能提供我們作為科學計算的基礎。因此，一切估計、檢定、信賴區間...，便有一套學理使它們能夠各安其位。

時至今日，抽樣仍是少數被公認為屬於正統統計的工作。它的主要功用是全面的：設法用少數幾個參數去了解整個母體的大略情狀。政府要做抽樣，其道理和我們要時常去做體檢一樣：要查證國家的健康。

執行一個大型抽樣是複雜而嚴謹的工作。最大的抽樣是全國性的普查，它所佔用的人力物力是驚人的。大型而經常性的抽樣，代表的是這個國家對於本身的認真程度。從技術面來說，因時、地、狀況的改變，時時需要將預設的抽樣計畫修補改訂，這裡面需要的技術性其實極為繁複。可惜的是：學院中的學者對於這一方面的工作，因為基本上插不上手而貢獻不多。

抽樣是以全面的角度來了解母體，是大官們的角度。但是，科技上的實際工作，則是需要尋找各因素之間是否有重要的關聯。而這要靠實驗設計。

## 農業實驗

一般以為，費雪 (R. A. Fisher, 1890 - 1962) 的工作，若不是打開了近代統計學的大門，也至少是具有繼往開來的氣勢和影響。他的書 “On the mathematical foundations of theoretical statistics” (1921) 雖談的是「數學基礎」，但是他工作的動機，卻是非常務實的農業實驗。真正的需求，不是來自小布爾喬亞式的「仕女品茶 (lady tasting tea)<sup>8</sup>」的休閒題材，也不是來自「身高和體重的關係」這類的習題型的應用 (Galton 1877)，而是來自農業實驗：我們需要有效地安排諸多因子和它們的水準，以期能使得每一個觀測值都能發揮出極大的效用。這樣才能學到怎樣的實驗安排或肥料配方可以增加小麥

---

<sup>7</sup>分水點在 Neyman (1934)。

<sup>8</sup>但這也許是「假設檢定」的濫觴。見 David Salsburg (2001)。

的產量。這些是早年的 know-how。早先的科技，主要是在農業。這些工作的主題所包含的，不止是科學上的求真，也充滿了實務上、經濟上的意義。我們知道，歐美各國的工業，早期根本沒有競爭。所以早年還談不上統計在工業上的應用。

這樣的客觀條件，產生了在 Rothamsted Experimental Station<sup>9</sup> 所發展出來的「實驗設計」。早期台灣的統計學，也以農業實驗為主（台大生統實驗室，成立於 1946）。早年台灣的經濟，靠農業很撐了一段時候<sup>10</sup>。今天台大的農藝系仍然是台灣生物統計的重鎮。

實驗設計的主題，談的是如何有效地蒐集數據，以避免結果混淆和增加實驗的精度。農業實驗，因為需要等一段時間才看得到結果（例如收成），先天上需要及早安排各種條件，同時執行，而比較不宜做每次只變動一個因子的安排。因此考慮的角度和傳統的工程師想法不同。這類實驗，著重的不在於突破性的發展，更在於大體上關鍵技術已知時的細部微調。因此雖然是以農業問題作為啓蒙，但對於今日的工業，也有重大的影響。

因為在設計一個實驗時，觀測值都是將來可能發生，但目前還沒有看見的值，無可避免地我們得用  $X, X_{ijk}$  等隨機變數來先做理論上的推導。這樣的架構基本上是整個頻率學派對一般統計推論的架構：所有可能的觀測值，在推導理論時，都被看成為隨機變數，而差不多只在最後一步，我們才把真實的值代入。這對於實驗設計而言，當然十分自然。

又因為要對將來要做的實驗做設計的工作，無可避免地我們要懂得：甚麼是好實驗，甚麼是差勁的實驗。因此諸如樣本數的平衡、因子混淆、分區... 這些概念，就需要用較數學的方式加以詮釋。這讓我們能認真地對待一些直觀，並適度地將它們凝聚成可以驗證的條件。

科學實驗的目的是求得知識。現代的科技，近五十年來的進展，大概是超越了過去一百年前再往前上溯五千年間的總和。主要的理由來自人類開始懂得如何有系統地來由實驗一步步地向真理逼近。

## 敘述統計

較早的統計，大多是描述某一母體的統計學 (descriptive statistics)。時至今日，我們仍然在學如何用少數幾個參數 (parameters) 來描述一個母體。「母體 (population)」的概念，也許是最要緊的統計想法。這是最早、最基礎的「模型」。進一步地，我們有所

---

<sup>9</sup>這據說是世界上最早的農業實驗站，也是英國最重要的農業實驗站 (1843 - )。但他們不知道，康熙曾做過水稻的品種改良工作，而康熙末年是 1722。

<sup>10</sup>在一些實驗農場裡，還能看見實體的 Latin square。

謂的「相關係數」。這才談到了兩個母體或者最簡單的直線迴歸。

母體的概念，既抽象又實際。自實體的角度來看，它是一個被賦予意義與結構的集合，例如某一個縣的所有公民。進一步看，則又包括了集合元素的一些如  $X_{ijk}$  的量測量。而統計學的想法則是設法賦予這些量一個靈魂：它們的分布或者聯合分布。

因此對於同一母體，我們就有好幾個角度來看。從一組隨機樣本來看，是古典的敘述統計。從它的分布來看，則是它的機率結構。從它的特徵函數 (characteristic function) 來看，則相當於把問題從現有的空間搬到頻率域 (frequency domain)。當然，我們還能從它的「經驗分布 (empirical distribution)」，來看

$$F_n(t) := \frac{1}{n} \sum_{i=1}^n I_{[X_i \leq t]} \rightarrow F(t)$$

這最基本的關連。這些基本的結果一個牽扯著另外一個，但又都關照著同樣的事：用不同的角度來審視一個母體——實務地、抽象地、解析地、機率地……。

## 古典和現代

從敘述到推論，中間僅隔了一線：但是古典和現代就在此分野。我們將問題中的數據架構在某一個機率模型裡，而忽然可用的解析工具就多了起來。我們學到的數理統計大概都在這樣的條件下誕生。它賦予古典的數據一個骨架，使得討論漸趨嚴謹精緻<sup>11</sup>，這相當於給原始的木偶一個生命，使得它可以活蹦亂跳，並且有時還說謊<sup>12</sup>。這當然同時也蔓生了許多枝節，有時也不免將我們導入歧途。

在某種意味之下，數理統計的架構，提供給我們一個理想的想像空間。它既能概略地描述數據所給出一般印象，又提供我們能夠精密計算的可能。在精神和實用上都是無可比擬並且很難拒絕的工具。現代的統計學，清一色地建在這樣的架構上。

統計學的基本理論架構——不論的假設檢定還是參數推估——是美麗而眩目的。但過度地強調細節，不免有如今日女仕們喜愛的名牌精品：雖然琳瑯滿目，但總覺得沒有真正到位或者處心積慮地想隱藏甚麼。精雕細鑿之後的統計方法，似乎仍然只能解最初的問題。畢竟，濃郁的化妝和高明的搭配，只能增加附加價值，和本質關係不大。科學要求的是：是否解決了本質不同的問題，是否看透了現象背後的真正意義<sup>13</sup>。至於補足

<sup>11</sup>這是學者們的最愛，因此生出許多論文來。

<sup>12</sup>所有的具都有可能被人謬用，統計自不例外。

<sup>13</sup>我們要尋找的是浣紗的西施，而不是在館娃宮裡的西施。前者是重要的新發現，後者只是儘力的推廣而已，學術的意味不深。

所有的漏洞, 固然不可輕忽, 但仍然是餘事。

## 基本問卷和迴歸

然後社會科學大行其道。或者, 我們可以說: 統計方法在社會科學裡大行其道。因為, 以量化為主要手段的討論, 是任何號稱是科學的學門最不能拒絕的誘惑。當然, 商業的行為更不能。量化需要數據, 而抽樣和問卷是社會科學中蒐集數據的基本方法。不能避免地, 問卷裡會包含了研究者認為「可能會發生影響」的自變數。因此基本問卷的型態, 幾乎是所有統計軟件的基本數據格式, 而迴歸變成標準的最起碼的分析工具。這些都可以放在傳統型的 Fisher 架構上來討論。並使得一些統計名詞, 例如  $p$ -值,  $R^2$ , 因為實在常常在相關的技術文件裡出現, 便逐漸演變成為現代科學語言的一部分。這一類的問題, 包括了絕大部分的各門各派所鑽研的種種模型。因為在某一個變數與其它變數中尋找關係, 是大部分研究者所能希望得到的最有力的結果。一般的標準迴歸及其它類似的迴歸的發展, 完全是基於這樣的需要。

時至今日, 軟體的發展已讓建模 (modeling) 的工作, 有了長足的進展。幾乎是所有已證明可用的模型 (或者略改幾個參數就可用的模型), 大概都有現成的軟體可用。基本上使用者只要網路蒐尋的技術夠好就行。

這是可喜的現象, 同時也是可憂的現象。因為這類模型背後的計算繁複, 基本上很難找尋其他公正有力的人士去再度獨立地來驗證這些電腦跑出來的結果是對還是錯。這是科學研究執行上的問題, 是技術上的, 遲早要解決。

## 承認隨機性

統計學發展主要的影響, 是科學界開始承認隨機性的存在, 並務實地和「誤差」共同生活。我們不止有了風險的概念, 並且還能進一步將它量化。

將這個概念推到商業的極致, 便成為今日的保險業。

承認隨機性並試圖針對它加以計算, 是新的想法。科學不能避免量測, 而量測不免誤差。我們經由誤差來了解模型, 經由變異來了解風險: 這背後是我們對於正態分布性質的澈底追尋。

這樣的進步是劃時代的。因此證券的投資比例 (portfolio), 可以用一個多維的 mean-variance 架構來加以表現, 這是所謂的資產定價模型 (capital asset pricing model, Sharpe (1964)); 在另一面, 衍生性金融商品的價格, 必然地和選擇權的對象的價格的

變異有關，這是所謂的 Black-Scholes (1973) 模型。這些都是精彩的論證。

這些理論的發展，是投資學上的大事。世上最多的、也是每天都在固定蒐集的數據是金融證券交易方面的數據。這也是資金極為密集的所在。能引起樣多的關注，甚至於好幾個諾貝爾經濟獎的授予，並不是偶然的事。

在另一方面，財務理論的精神，在這樣的理論架構之下，也得到尊重和保證。例如衍生性商品的定價，基本上還是要靠「在有效率的市場上，應該沒有任何套利的機會」這樣的前題，來加以詮註。

## 品質管制的興起

戰爭除了破壞之外，還大量地刺激了消費和需求。二次大戰 (1931 – 45) 使得大規模而快速的軍火生產，不只成為必要，且延續到所有和軍品相關的工業。這時便自然地產生了諸如品質管制 (quality control, QC), 運籌學 (operation research, OR) 這些學問。這裡面，QC 的部分，並沒有太多的正規統計推論。但是卻是有聲有色的應用。而 OR 卻較偏向最優化的數學工作。但這兩者，不論是規模和影響都十分巨大。比較早期的 OR 專題之一是排隊的理論。這也是來自工業上的需求：電話網路的安排 (趙民德 1975)。

早先關於 QC 的數據，是生產線上自然產生的 (production line data)，我們只要費心加以測量便好<sup>14</sup>。QC 的技術，著重在快速有效，並且還要有在第一線的人員就能明白的意義。雖然，在技術上，如果我們能充分懂得二項和正態分布，早年的 QC 大概就出不出這樣的範圍。但這不是我們的重點。我們要強調的是：數學上困難度和統計學是否有重要的影響無關。課題本身的專業意義，要重要得多。

OR 的另一個方向，又漸漸演變而為管理——數量管理。這些不全是我們所熟知的統計學，但是不論是簡易的「品管七大手法」、TQM 或者目前還在大行其道的 6- $\sigma$  系統，其基本可以實際運作的部分，仍然是最起碼的統計<sup>15</sup>。統計方法不斷以不同的型態出現，例如在資料倉儲 (data warehouse) 裡的核心，是一個稱為 cube 的數據結構，而這只是我們常用的多維列聯表。

到目前為止，數據的形式都還是完整的。當我們說「看到  $X$ 」，意思是說，我們可以完全看到  $X$ 。但是，漸漸就出現了「只看見一部分  $X$ 」的情形。我們下面所談的就是這樣的情形。

---

<sup>14</sup>現在則因為有很多自動量測的系統，很容易附加在生產線上的各重要環結處，因此這類數據的量極大。

<sup>15</sup>別的要靠領導者的決心。

## 設限資料

我們先是知道資料有用，並且盡力蒐尋，努力將其中的資訊擷取展示。早先的資料都是完整的資料，對於不完整的資料的分析，一方面需求不高，另一方面我們的能力有限（例如五零年代的電腦，不能和現代相比）。因此早年的發展，除了一二個「能算出解析解」的特例外。結果較少。

不完整的資料，來自實務。對於純粹的「遺失」，我們不妨將它們簡單地看成「根本沒有蒐集」。例如我們原想取 100 個數據，但其中某一個沒有取到，那麼我們就不妨將數據看成一組有 99 個數字的隨機樣本。—— 這個想法，基本上是對的<sup>16</sup>！但是，若是我們對某些數字「只看到一部分」時，又應該如何處理<sup>17</sup>？

「只看到一部分」，可能是時間不夠，可能是目標已達。這些都是簡單的「設限數據」(censored data)。例如對於癌症的臨床研究，通常不會超過五年。時間到了，如果病人還健在，我們只知道他「存活了至少五年」，但看不到他會存活多久<sup>18</sup>。這是第一類的設限 (type I censor)。這在工業可靠度的研究中也時常發生：現在的工業產品愈做愈好，在 100 個測試的產品裡，或者只有 70 個或更少的真正是損壞的。那麼，餘下的 30 個未壞的產品，難道不包含可用的資訊？

在生物臨床實驗裡，不完整的資料還會以其他形式出現。例如病人可能是死於車禍。由於「生命分布」愈來愈長，關於正確處理種種設限數據的技術，才會風起雲湧地被發展出來。這背後的力量是需求：製藥的確認以及醫學上的求真。當然它們都需要相當的數學技巧和深度，但那不是重點：重點是，有太多這一類的數據需要我們去分析。所以才發展出新的統計方法。

## 加速壽命測試

對於人體，我們不能對他做出過於冒犯的事。但是，對於電子元件，我們大可以加上一些條件，使它們會「壞得快些」。例如增加測試時的溫度、電流強度等。這在可靠度的研究和實務裡，叫做「加速壽命測試」，是臨床醫學裡沒有的一項。

對於某些產品，它基本上不會壞，那麼又該如何做？例如發光二極體 (LED) 的壽

---

<sup>16</sup>但這在兼顧到各變數間的平衡時，就有了問題。這產生了種種的插補法，而這在抽樣資料的分析時常被用到。

<sup>17</sup>簡單的方法是將它想成遺失。很多人現在仍是這樣做，但這樣會數據裡的損失資訊。

<sup>18</sup>存活分析，在醫學上是和臨床實驗相關的。農業只為了吃飽或吃好，工業則多是為了衣、住、行、樂，而近代的生物醫學則是著重在活得長。這裡面有各國政府的大力支持，是統計學得以在此大幅活動。

命，理論上是十萬小時。沒有一個工廠會等那麼久，因為一萬小時都嫌太長。這時，我們就要考慮「衰變分析」，例如每周測一次它的光度，看光度減弱的程度來研判估算該產品的可靠度。

高可靠度分析，來自精密工業的要求。但是，科技的發展又讓我們有其它不同型態的數據。在在需要統計學家的投入。

## 數字型態以外的數據

數字型態以外的數據，比如晶元圖 (wafer map)，生物晶片 (micro array)，衛星遙測資料，腦部 X 光的橫斷面影像、煉鋼爐或者某一個化學反應槽的內部反應<sup>19</sup>。每一種資料，都有它蒐集的目的，也有它們想問的問題。但它們多半不是  $H_0$  vs.  $H_1$  的教本型態，我們也最好不要先三不管將它們化為假設檢定或者信賴區間。

但是統計學的基本道理，也許會有所修正，但不會有太大的改變。這都拜近四五十年年統計學者對於統計學基本認知的努力。但問題會一變再變，數據也會有不同的形式，只有不斷地從新型數據裡來一再深探，統計學才能夠真的成長。

很多統計學者是學數學出身。數學裡有很多有名的大猜測 (conjecture)。任何一個默默無名的小數學家，若能證明其中之一 (甚或一半)，都會立刻成名。這是數學家看重的。但是，統計學裡沒有甚麼有名的重大猜測。你能證固好，根本不去想它們也不會被別人說功力不夠。因為若是我們去看真的問題和真的數據，我們根本沒有時間去玩數學遊戲。

## 回溯

統計學的發展，是跟著數據的形態和問題的本質來改變的，不是因為我們會做它背後的數學而發展的。但在統計研究的工作的過程裡，不時會有一些還算是有趣的小數學問題出現。這是一個陷阱。不要捨本逐末，輕易地就掉在數學的陷阱裡。賀吉士 (J. L. Hodges, Jr., 1922 - 2000) 曾說過：不要因為 (統計的) 問題困難而去做它；也不要因為它難而不做。研究的選題，主要還是要看問題值不值得<sup>20</sup>。

## 統計科學還是統計工業

---

<sup>19</sup>對這類的數據，統計學開發新方法的速度常跟不上硬體的發展，因此也給我們一些空間。

<sup>20</sup>這是我在 1963 做研究生上課時學到的，對我一直有很大的影響。

前面我們所提到的多半是統計學，或者更進一步的統計科學。但是，統計科學或者數據科學還是有一點不夠，我想看到的是統計工業。

工業意味著不斷改進的、有市場的產品和日益增加的相關就業人口。如果統計只是在大學裡或者研究單位裡面存在的話，那麼個行業的蕭條，也是可以預期的。

工業的另一個意涵是較難存在著單兵作戰。關張趙之流的勇將，在攻城掠地之時，還是需要大部隊的支援。工業化的結果，除了世俗化之外，還暗示著有更高的附加價值<sup>21</sup>。有了小兵，才襯得出大將的不凡，而小兵也需要養家活口，盡他們的社會責任。

我曾在化工廠裡看過滿牆的螢幕：每一個都在監測某一重要生產環節的狀況和反應。我聽過虛擬的晶圓工廠，我也看到企業裡開始有「資料價值發展部」的編制。教育測量服務 (Education Testing Service, ETS) 所做的，除了 TOFEL 之外，還有不斷改進中的測驗技術。這是將 item response theory (IRT) 充分發揮的工業。ISR<sup>22</sup> 和 NORC<sup>23</sup> 基本上都是法人化之後的調查機構。而 Morgan-Stanley 難道不是在蒐集並發布數據？那些作債信評等的大公司，處理的不是數據？甚至於華爾街日報也做的是類似的工作。——他們只是將某些統計工作專業化，不斷地加上新的價值罷了。

## 結語

魚要有水，統計要數據。水若污染，魚會死，因為它改變自己不夠快。統計要能存活，或者，一個統計人想要存活，不要只守著一畝三分地，只做自己或者別人方法論的推廣，更不能甚麼問題都套上同樣的三斧頭。要不斷做新的投資。

自從人類懂得結繩記事，我們就開始有數據。科學的發展，使得現在更是一個充滿數據的時代。蒐集它們都有一點目的，有些數據來之不易，成本甚高（例如航太遙測的數據），它們背後的目的更大。數據是時代的大河，千里而來，出海而去，我們知道泛濫後的土更為肥沃。

統計學的長遠發展，要建立在這樣的基礎上。

這是一個系列講稿的開場白。統計雖然無可避免地用到很多數學，但不是數學。數學不夠好，做統計的工作，不論是研究還是實務，都很辛若。但是數學太好，其實也有害。——每一個學門，都有她自己的味道。

---

<sup>21</sup>這時候就有了種種包裝。

<sup>22</sup>Institute for Social Research, University of Michigan.

<sup>23</sup>National Organization for Research, University of Chicago.

統計學的味道，來自她的數據。因為這是她的本質。

## 參考資料

- 趙民德 (1975). 從鼠牙談起——五花八門的電話研究工作。科學月刊全文資料庫：  
<http://140.111.102.168/science/content/1975/00110071/0008.htm>
- Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637-654.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, A* **222**, 309-368.
- Galton, F. (1877). Typical laws of heredity. *Nature* **15**, 495-533.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Roy. Statistical Society, B* **97**, 558-606.
- Salsburg, D. (2001). *The Lady Tasting Tea*. W. H. Freedman and Company.
- Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* **19**, 425-442.
- Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press.

file=D:/my stat articles/paper-7.ctx, printed February 5, 2007