part ONE 第一部

當下思維

Current thinking

第2章 無中生有的觀點？

許多大數據的愛好者最愛宣稱的說法是規模至上。換句話說，隨著資料量倍數成長，再加上我們傳輸、儲存、分析技術的進步，我們正以過往、無法完成之創新能力，進行探勘並瞭解有價值的信息。據稱，現在我們不再依賴傳統的統計隨機抽樣，以代表性樣本做資訊推論，而是可以根據所有的資料做判斷。

麥爾荀伯格和庫基耶指出[1]，長久以來我們使用隨機抽樣，是因為它可以把大規模的資料蒐集所衍生的作業困難，轉化成可管理的流程。但在大數據時代來臨之際，有人主張隨機抽樣的觀念已經過時，是「退而求其次」的次要選項。當擁有龐大的母體資料時，又何必使用樣本呢？

大數據確實為企業創造了很多獲利的新商機，但大數據的出現也伴隨了很多誇大的說法，例如宣稱這些資料都是客觀的、都很可靠。有些人主張大數據超越了科學的傳統局限，本章將探索這些說法。

你鎖定的目標對象是誰？

關於資料蒐集，有一個存在已久的迷思：樣本愈大愈有代表性。1936 年 10 月，就在當年的美國總統大選之前，《文學文摘》（The Literary Digest）所公布的民調就是著名的例子[2]。他們寄出 1000 萬份問卷給選民，收到約 230 萬份的回函。那些民調的對象是取自雜誌的訂戶清單、汽車登記名單、電話名單、俱樂部會員的名單。

有趣的是，該雜誌在前四次總統大選中也做過類似的民調，都成功預測了選舉結果。不過，這次他們預測共和黨的挑戰者阿爾夫．蘭登（Alf Landon）將打敗現任的民主黨總統羅斯福，結果錯得離譜。羅斯福不僅沒輸，還贏得壓倒性的勝利，那家雜誌社也在兩年後關門大吉。

事後檢討分析顯示一些值得借鏡之重大失誤的發生，其中最主要的因素是當時的金融環境，美國正陷入有史以來最嚴重的經濟蕭條，那些從（昂貴的）雜誌訂戶清單，以及汽車、電話、俱樂部會員的名單中抽取的受訪對象，自然也包含較多財力優渥的人士，他們的政治立場比較偏向共和黨的候選人。在以往的選舉經驗中，所得水平差異不會產生嚴重的統計推認偏誤，但是在經濟大蕭條時，那絕對是很大的問題。

另外一個問題來自於所謂自我選擇的現象。係指願意花時間填寫回覆問卷的人，很可能和不想花時間填問卷的人有不同的投票傾向。

樣本的偏誤來源

當然，在研究抽樣設計時，需考慮到偏誤問題，過往之長期經驗累積也幫助我們瞭解如何有效管理及降低偏誤發生。我們簡略來看一下可能出現的偏誤類型。

1. **自我選擇偏誤**。當個人主動加入群體時，就會發生這種情況，因為主動加入的個體可能和研究人員想分析的目標母體明顯不同。選擇寄回《文學文摘》問卷的人就有明顯的自我選擇成分。
2. **涵蓋不全偏誤**。當母體內部分群體遭到忽略排除時，就會發生這種情況。例如，《文學文摘》未包含財力較差的個體，那些人比較可能支持現任的民主黨總統羅斯福。
3. **存活者偏誤**。鎖定某些經歷下倖存之人或物，無意間就忽略了沒有存活下來的人或物，就會發生這種情況。例如，沒有把已經倒閉的公司納入績效研究中，因為他們已不復存在。

除了挑選的樣本以外，其他的偏誤來源也值得一提。例如，受訪者不願透露實情（飲酒習慣調查就是一例）、回覆率低、調查中的用字遣詞／問題順序等等。

誠如古德（Good）和哈丁（Hardin）所言[3]：只要多努力，就可以解決多數的取樣問題。

> 透過仔細與長期的規劃，可以減少或消除許多可能的偏誤來源，但是要做到消除所有的偏誤，那種情況少之又少。接受偏誤是無可避免的，接著就努力去找出那些遺漏的例外，提報出來。

抽樣的優點

顯然，普查法（亦即可取得整個母體的資訊來做分析）對研究人員很有吸引力。但是，實務上，依然以抽樣法為主，主要原因如下：

- 處理成本：許多品牌擁有龐大的交易資料庫。他們通常會剔除約 10%的記錄，不然分析處理時間和成本都太大了。品牌想確保他們握有足夠且允當、具代表性之資料集，可以深入探索特定的族群或區隔之行為。
- 品質：統計學家愛德華茲．戴明（W Edwards Deming）的研究對品質衡量管理有重大的影響，他主張抽樣法的研究品質往往優於普查法[4]：「相對於全面涵蓋的情況，取樣可讓研究人員做更好的訪談（測試），更徹底地調查缺失、錯誤或可疑的資訊，做更好的監督和處理。」研究結果也證實這樣說法。一項研究顯示，90%以上的調查錯誤是來自非抽樣錯誤，只有不到10%是來自抽樣錯誤[5]。
- 速度：取樣通常比整個資料集更快提供相關的資訊，因為蒐集、處理、分析資料集的流程可能很耗時。

樣本愈大不見得愈好

市調人員都很熟悉一個重要的議題：隨著樣本的規模變大，誤差幅度會跟著縮小。不過，應該注意的是，這個說法無法無限擴大延伸：在樣本規模介於 200 到 1500 之間時，隨著樣本數變大，誤差幅度縮小得明顯，但是樣本數再持續大下去，其效果就接近持平了，如表 2.1 所示。

所以樣本愈大雖然可以改善精確度，但改善的幅度很快就縮小了。此外，每次選取子樣本時都需要重新校準誤差幅度，所以市調人員一開始通常會盡可能擴大取樣範圍。

表 2.1　誤差幅度和樣本規模的關係

樣本規模	誤差幅度
200	±6.9%
400	±4.9%
700	±3.7%
1,000	±3.1%
1,200	±2.8%
1,500	±2.5%
2,000	±2.2%
3,000	±1.8%
4,000	±1.5%
5,000	±1.4%

大數據和抽樣

大數據常伴隨的根本假設是：所有的記錄都可以取得，所以我們處理的是母體，不是樣本。麥爾荀伯格和庫基耶認為掌握全部的資料集是有助益的，因為：

- 不僅探索資料時更自由，研究人員也可以深入探索以前難以想

像的細節。

- 這種資料蒐集的方式比較不會出現抽樣偏誤。
- 有助於發現之前潛藏的資訊，因為資料的規模讓研究人員可以看到小樣本中看不見的資料關連性。

他們舉 Google 流感趨勢調查（Google Flu Trends）為例。Google 運用匯總的搜尋詞彙來估算流感活動，分析甚至可以顯示流感蔓延到各城市的程度（我們在第六章會深入探討它的效用）。網絡理論科學的頂尖研究者艾伯特－拉茲洛・巴拉巴西（Albert-Laszlo Barabasi）所做的研究是另一個例子。他從一家無線電信業者取得匿名手機用戶的四個月使用記錄，該電信商服務近五分之一的歐洲人口。巴拉巴西和團隊運用「每個人」的資料集，發掘多種關於人類行為的見解，他認為使用較小的樣本可能無法達到那樣的效果[6]。

這種想法確實顯示，大數據提供想瞭解人類行為的研究人員夢寐以求的素材，但是大數據的使用還涉及一些挑戰，我們討論如下。

大數據取樣

「不要使用全部的資料」，這種概念乍聽之下悖離直覺，卻是務實的做法。2010 年教育和政策研究機構亞斯本學院（Aspen Institute）在大數據開始掀起熱潮時，發表了一篇報告[7]，該文提出一個問題：「資料更多，會不會反而效果更差？」

該文引用 Google 首席經濟學家范里安的說法，討論「較小的資料集是否無法成為大數據的可靠代表」這個假定命題：

> 在 Google，工程師從每日資料中取 1/3% 作為樣本，他們用這些代表性樣本來計算統計數據。你從隨機樣本得到的結果，通常和全部資料看到的結果一樣好。

你鎖定的對象是誰？

大數據的支持者主張，大量資料不僅可以讓你找到你不知道要找的東西，也可以衍生實用的新見解。只要你知道該問什麼問題，你確實可以找到有意義的答案。此外，此類說法也需要確定一點：你拿到的大數據確實代表你感興趣的整個母體，而且其來源有代表性和準確性。

我們回到巴拉巴西和無線電信商合作的研究。那些資料很可能代表數百萬的個體，但是在做這種概括的假設以前，你必須更清楚瞭解那是怎樣的電信業者，才能掌握那個情境和環境。例如，它的用戶是否以商業用戶居多？如果是的話，那適合拿來研究嗎？或者，它的用戶是否年齡偏大或家庭導向？瞭解這些之後，你才能開始判斷可能出現哪種偏誤。

麻省理工學院公民媒體中心（MIT Center for Civic Media）的凱特・克勞馥（Kate Crawford）不確定大數據總是比較好[8]。

> 資料和資料集並非客觀的，而是人類設計出來的。我們賦予數字表達的能力，並從中推斷結論，透過我們的詮釋界定其意義。蒐集與分析階段的潛藏偏誤都是很大的風險。對大數據的運用來說，那些偏誤和數字本身一樣重要。