論述 》專論 · 評述

數據分析的美麗陷阱：倖存者偏誤

在 21 世紀，機器學習、大數據和人工智慧依次崛起，使數據分析成為解決複雜問題和制定有效策略不可或缺的工具。本文透過案例說明，帶領讀者了解數據分析可能產生的偏誤，避免被數據誤導，進而提升分析與洞見萃取能力，除讓數據能說話，更能讓數據說對的話。同時亦強調數據分析並非萬能，惟有搭配專業團隊的領域知識和經驗，才能讓數據分析發揮價值。

孔令傑（國立臺灣大學資訊管理學系副教授）

壹、前言

數據分析是一種透過統計、數學和電腦科學等方法，對收集來的數據進行解釋、歸納和預測的過程。這項工作不僅僅是對數據的處理，更是一種深度思辨的過程，須結合領域知識、合理推斷和邏輯分析，以確保結論的準確性和可靠性，也因此，數據分析的過程中往往伴隨著種種挑戰，如數據錯誤、偏見以及對分析結果的不當解讀。在大數據的時代，擁有數據固然重要，該如何正確運用數據更是一門學問。本文將透過案例分享，探討數據分析中常見的錯誤，並提供相對應的解析和操作心法，希望讀者們在擁有數據、分析數據的同時，也能擁有寶貴的洞察力，更能夠避免被數據誤導，讓數據分析能真正地幫助到組織與自己。

貳、案例說明

一、健檢中心的臨時取消率分析

許多人做數據分析時，容易被片面的數據或圖表誤導。讓我們以一組從眞實場景虛構出的數據來給讀者們一點小小的挑戰。

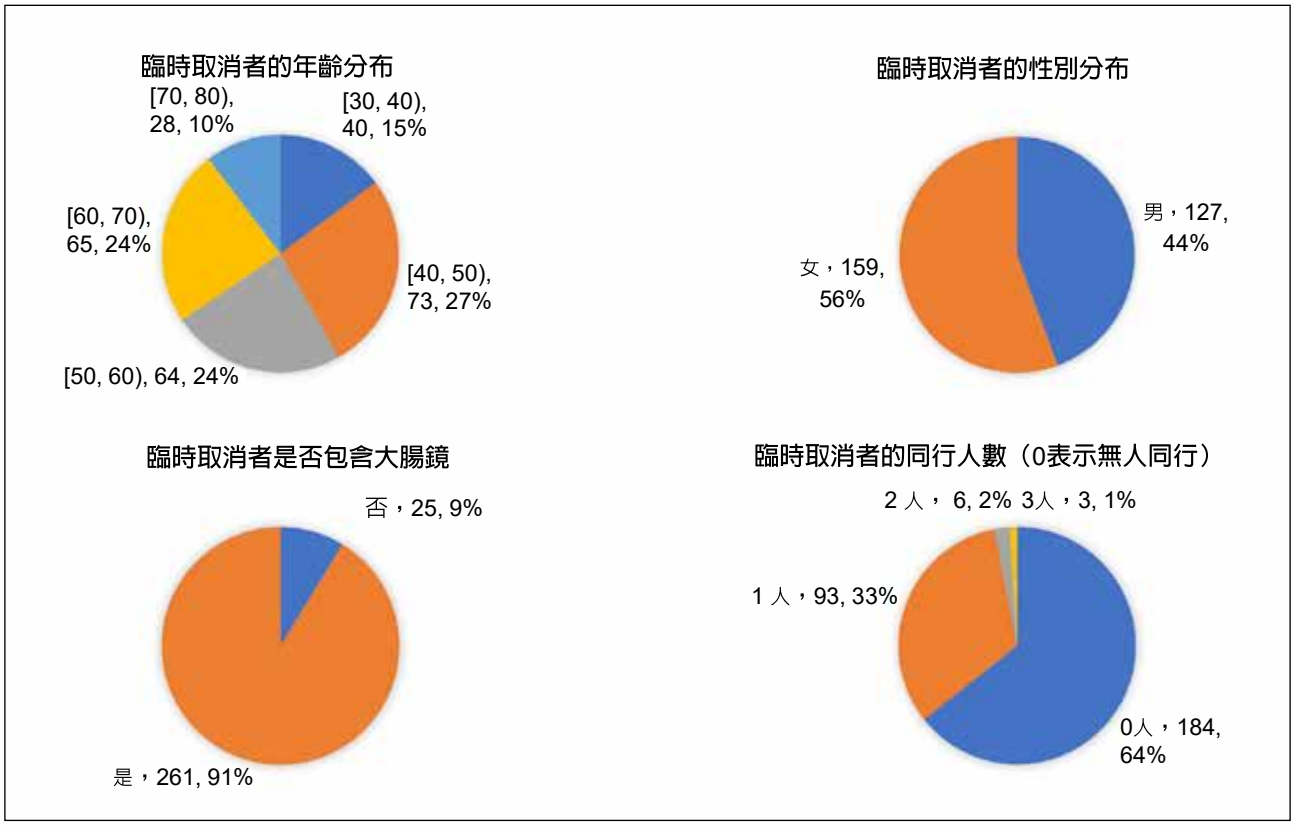
想像您在一家健康檢查中心服務，該中心經常遇到消費者預約健檢後卻臨時取消（在健檢當天或前幾天取消），導致珍貴的名額被浪

費，一來傷害中心的營收，二來也讓中心無法服務到儘量多的民眾。爲了減輕這個問題，中心請您分析過往資料，看看比較容易臨時取消的都是怎樣的人，或許以後可以針對這類消費者多做提醒或甚至加收保證金等等。

爲此，中心從過往某一年的所有預約中隨機抽出部分一般民眾的預約紀錄，共 8,431 筆，其中 286 筆臨時取消，比例約 3.4%。針對每筆預約，我們有數個變數，包含健檢者年齡、健檢者性別、該次預約是否有包含大腸鏡、臨時取消者的同行人數等。您的同事先做初步分析如附圖，呈現在您的面前。看了附圖，請問您覺得比較容易臨時取消的都是怎樣的人？以後若要多加提醒，應該提醒哪些人？

如果您剛剛認眞地解讀了附圖，可能馬上看到「有做大腸鏡」、「女性」、「獨自前往」、「40 幾歲」這幾個區塊，並且心想未來遇到這類預約時，可能要特別留意。但！再仔細多想一會兒，您應該可以發現，附圖的分析事實上很

附圖　健檢中心預約臨時取消之初步分析



資料來源：作者自行繪製。

可能是誤導人的。以有無大腸鏡爲例，附圖只呈現出「臨時取消的人之中有 91％有做大腸鏡」，但我們並不知道在所有人中有多少比率有做大腸鏡，搞不好是 95％呢？舉個更誇張的例子，要是有天我們看到「臺北市統計，去年騎機車車禍受傷的民衆中，有 99％都有戴安全帽」，也不能說戴安全帽的人比較容易車禍受傷，畢竟大概 99.99％的機車騎士都是有戴安全帽的。

若我們更完整地分析數據，以大腸鏡爲例，總計有 7,861 人預約大腸鏡，其中臨時取消者爲 261 筆，換算成臨時取消率爲 3.3％；相較之下，未預約大腸鏡共 570 筆，最後臨時取消 25 筆，臨時取消率爲 4.4％。換言之，有預約大腸鏡的，其實是比較不容易臨時取消！仔細想想也不奇怪，畢竟如果要做大腸鏡檢查，一般都要提前三天進行低渣飲食，當天早上還要喝瀉藥清腸，都這樣辛苦地準備了，一般人應該是非不得已，不然都會千方百計地完成檢查，以免過一陣子又要再來一次。又以結伴同行爲例，事實上單獨預約檢查者共有 5,895 人，其中臨時取消 184 筆，臨時取消率爲 3.1％，而有結伴同行的 2,536 筆則有 102 筆臨時取消，相當於 4％的臨時取消率。換言之，獨自前往的民衆也是比較不容易臨時取消的。

從案例一可以看到，附圖的分析若要說是有哪邊有誤，那可以說就是「漏了分母」：只針對臨時取消的資料進行分析，而沒有分析作爲分母的全體資料。像這樣只分析「通過某個門檻」的部分資料，而因此對事實眞相產生錯誤理解，被稱爲「倖存者偏誤」。讓我們再多看一個例子，以對此有更多理解。

二、轟炸機該補強哪裡

「倖存者偏誤」被廣人爲知，可能要歸功於二次大戰期間英國軍方的故事。當時的盟軍會派遣轟炸機前往德國領土進行轟炸，經常被德軍砲火擊落，軍方因此希望能在轟炸機上補強鋼板以減少傷亡率。由於鋼板很重，不能用鋼板覆蓋整架飛機，因此軍方希望做關鍵重點補強。要怎麼知道哪裡是「關鍵重點」呢？軍方分析了執行轟炸任務後的轟炸機，發現多數彈孔分布在機翼和機尾，駕駛艙、發動機和油箱則很少被射中，於是他們決定「應該加強機翼與機尾的防護，因爲這是最容易被擊中的位置」。

對於前述結論，亞伯拉罕 · 沃德教授卻有不同觀點。他認爲這個研究的樣本只包含安全返航的轟炸機，而不包含那些因敵火射擊而墜毀的。反觀駕駛艙、發動機和油箱並不是不容易被射中，而一旦被射中就很難安全返航，這表示這些地方才是「關鍵重點」。經過討論，軍方最終採納教授的意見，增加對「幾乎沒有彈孔」的部位（駕駛艙、發動機、油箱）的防護，大幅降低戰機傷亡率。

參、數據分析的能與不能

前述兩個例子都讓我們看

到了「倖存者偏誤」可能會讓決策者對事實有錯誤的理解，進而做出錯誤的決策。有趣的是，這兩個例子也有著截然不同的地方。在健檢臨時取消的案例中，是有辦法得到事實眞相的：只要記得要將分母（全體預約紀錄）納入分析，就可以得到「沒有預約大腸鏡的消費者有比較高的臨時取消率」等等的事實眞相[1]。但在轟炸機的案例中，是沒有辦法得到事實眞相的，除非德軍允許英軍到德國領土調查每一架被擊落的轟炸機，不然根本不可能證明沃德教授的「駕駛艙、發動機和油箱並非不容易被射中，而是一旦被射中就很難安全返航」論點是正確的。換言之，在轟炸機的例子裡數據是不足的，分析者大概永遠無法「用數據證明」教授的觀點，即使那是對的。

那爲什麼英國軍方最後願意採信教授的論點，讓他們能做出正確的決策呢？故事中通常沒有詳細記載，但筆者認爲，軍方的領域知識（domain knowledge）應該扮演了重要的角色。軍方中顯然不乏飛行專家與戰爭專家，他們的知識和經驗多半告訴他們，駕駛艙、發動機和油箱確實並非不容易被射中，而且這些地方一旦被射中，確實就很難安全返航。也就是說，軍方雖然有被沃德教授提醒，不要落入錯誤數據分析的陷阱，但軍方並不是被沃德教授用數據說服並接受其觀點；軍方事實上是用自己的領域知識做了判斷，最終再輔以一點數據佐證。

肆、結語

近年來，許多組織和企業都積極收集、累積數據，數據分析的技術也持續蓬勃發展，這些固然是好事，但確實也讓部分分析人員誤以爲數據分析是萬能的，或者就算不是萬能的，也是唯一可信的，而忽略了領域知識與經驗的重要。數據分析非常有用，但領域知識和經驗也同樣有用。一個專業的分析團隊，固然不能只相信領域知識和經驗，但更不能只重視數據分析；唯有同時善用兩者並且找到好的平衡，才能眞正透過數據分析爲組織帶來價值。

註釋

1. 有些讀者可能有豐富的數據分析知識與經驗，已經想到「相關不等於因果」、「統計顯著性」等進階議題，這些確實都很重要，但本文受限於篇幅在此不進一步討論。❖