數據分析的美麗陷阱：淺談倖存者偏誤

文．圖／孔令傑

數據分析是當今的顯學，但也伴隨著種種挑戰，如數據錯誤、偏見以及對分析結果的不當解讀。透過本文，我們將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。

案例一：健檢中心的臨時取消率分析

許多人在做數據分析時，容易被片面的數據或圖表誤導。讓我們以一組從真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盡量多的民眾。為了減輕這個問題，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。

為此，中心從過往某一年的所有預約中隨機抽出了部分一般民眾的預約記錄，共有8431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。您的同事先做了初步分析如圖1，呈現在您的面前。看了圖1，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？

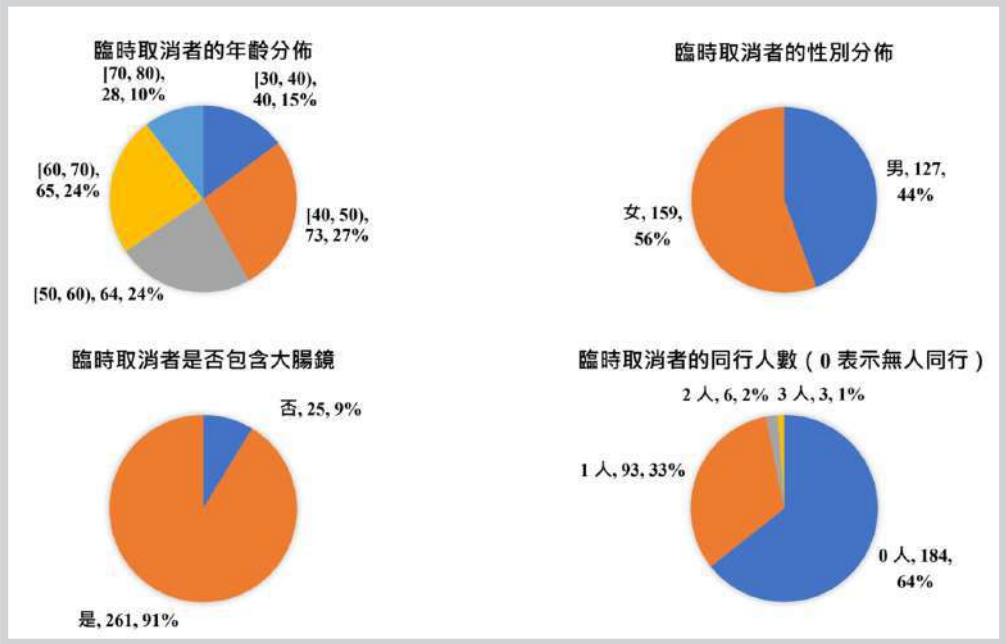

圖1：健檢中心預約臨時取消之初步分析

如果您剛剛認真地解讀了圖1，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，圖1的分析事實上很可能是誤導人的。以有無大腸鏡為例，圖1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？

若我們更完整地分析數據，以大腸鏡為例，總計有7861人預約了大腸鏡，其中臨時取消者為261筆，換算成臨時取消率為3.3%；相較之下，未預約大腸鏡的預約共570筆，最後臨時取消25筆，臨時取消率為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。

從案例一我們可以看到，圖1的分析若要說是有哪邊有誤，那可以說就是「漏了分母」：只分析做為臨時取消率分子的臨時取消資料，而沒有分析作為分母的全體資料。像這樣只分析「通過某個門檻」的部分資料而因此對事實真相產生錯誤理解，被稱為「倖存者偏誤」。

案例二：轟炸機該補強哪裡

「倖存者偏誤」被廣人為知，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，發現多數彈孔分佈在機翼和機尾，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，因為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。他認為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，而是一旦被射中就很難安全返航了，這表示這些地方才是「關鍵重點」。經過討論，軍方最終採納了教授的意見，增加對「幾乎沒有彈孔」的部位（駕駛艙、發動機、油箱）的防護，大幅降低了戰機傷亡率。

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，

圖 2：數據分析必須立基於領域知識與經驗，否則容易出現「倖存者偏誤」。圖取自 Unsplash by Carlos Muza。

是有辦法得到事實真相的：只要記得要將分母（全體預約紀錄）納入分析，就可以得到「有預約大腸鏡的消費者有比較高的臨時取消率」等等的事實真相[註]。但在轟炸機的案例中，是沒有辦法得到事實真相的：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論點正確。換言之，軍方雖然有被教授提醒了不要落入錯誤數據分析的陷阱，但軍方並不是被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（domain knowledge）做了判斷，知識和經驗告訴他們，駕駛艙、發動機和油箱一旦被射中，確實就很難安全返航了。

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忽略了領域知識與經驗的重要。筆者認為，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，才能真正透過數據分析為組織帶來價值。（本專欄策畫／資訊管理學系蔡益坤教授）

[註] 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。

孔令傑 小檔案

孔令傑副教授為國立臺灣大學資訊管理學學士與碩士、美國加州大學柏克萊分校工業工程與作業研究博士，返臺後任教於母校資訊管理學系至今。其研究方向以多邊平臺、數位經濟、供應鏈與作業管理為主。孔副教授除了教授「程式設計」、「商管程式設計」、「資料庫管理」、「資訊經濟與賽局理論」、「作業研究」等一般生科目外，亦在臺大 EMBA、PMBA、GMBA、EiMBA、管碩學分班等學程授課；曾獲臺灣大學教學傑出獎、教學優良獎、校內服務優良獎、優良導師獎、優良服務學習導師獎。其有多門數位課程上線於國際 MOOCs 平臺 Coursera，至今累積超過 90,000 名學生註冊修課。此刻孔副教授亦兼任校內進修推廣學院副院長、PMBA 學程主任，以及教務處數位學習中心副主任。