統計知識的應用

主要資料來源：《統計與真理》，C. R. Rao 著(美)，科學出版社，2004。

一般人可以利用統計知識在日常生活中做出各種決策、制定將來的計劃、決定是否投資或購買哪支股票。為了避免被五花八門的宣傳廣告迷惑，人們有必要掌握一定的統計知識。對一個國家的政府來說，統計學是一種為達到特定經濟目的和社會目的，以制定長期和短期計劃的工具。高深的統計技術可用於做出人口預測以及商品消費和流通需求的預測，也可用於制定經濟規劃。

在科學研究中，設計實驗計劃來收集數據、作假設檢定、未知參數的估計及對結果的解釋都離不開統計知識。在工業生產中，統計技術可用來維持和改良產品品質和增加產量。在醫藥開發和研究中，統計可幫助藥效鑑定及臨床試驗，以作出正確決策。甚至在文學領域也可用來鑑定爭議作者、法庭上可作為補充證據……，統計也都成了不可缺少的工具。

以下搜集了一些有趣的統計應用具體案例：

- 1985 年 11 月 4 日研究莎士比亞的學者泰勒，從 1775 年就保存在 Bodelian 圖書館的 9 節新詩，新詩只有 429 個字，沒有記載誰是詩的作者，這首詩會是莎比亞的作品嗎？1987 年二位統計學者 Thisted 和 Efrun 利用統計方法研究了這個問題，得到的結論是這首詩用詞風格確實與莎士比亞的風格非常一致(本書原著有詳細說明分析方法的細節，此處略)。
  (上海復旦大學的李賢平教授也利用類似的統計方法研《紅樓夢》的作者問題。因為歷來認為前 80 回為曹雪芹原著，後 80 回為高鶚所續。他的結論是：前 80 回與後 80 回確實出自兩個不同手筆的作者，但是，中間諸多章節至少經過了 5、6 個人修改過。)

- Yardi 於 1946 年在沒有任何訊息的情況下，利用純度量化方法對劇本求出以下頻率：(1) 冗長的最後音節； (2) 完全的分行； (3) 帶有終止符，但沒有分開的行； (4) 對話文的總數； 將莎士比亞文學作品的風格變化加以量化，再利用他對其他記載了出版年月的作品，推斷漏掉出版年月二本喜劇的發表時間。其中《錯誤的喜劇》大約是在 1591~1592 年冬發表，而《愛的徒勞》發表時間大約是 1591~1592 年春。

- 柏拉圖有 35 篇對話、6 篇短文和 13 封信的作品並沒有寫作時間年表。

Boneva 於 1971 年基於每一作品中最後 5 個音節的 32 個可能特徵的頻數分布，求出相似性指數為「定性終止」，並假設寫作時間相近的作品風格相似，推斷出了柏拉圖那些作品的時間年表。

- 「語言年代學」是利用語言之間相似性的大量訊息和複雜推理，語言學家能夠鑑定語言的一些主要流派，如印-歐語言樹、馬來-波尼西亞語言樹、印-藏語言樹……等等，用來歸類不同語言之間的語系關係。

- 20 世紀早期，哥本哈根卡爾堡實驗室的 J. Schmidt 發現，不同地區所捕獲同種魚類的脊椎骨和鰓線的數量有很大的不同，而在世界各地海域捕獲的鰻魚樣本，發現幾乎都有一樣的平均值和標準差，由此，他推斷鰻魚一定是由「某個」公共場所繁殖的，後來，科考船果然發現了這個地方。

- 丹麥遺傳學家 W. Johannsen 進行了一項實驗，他取了大量的豆子，秤它們的重量，由這些重量做成直方圖並擬合了常態分佈的曲線。然後，他從中取出大的和小的豆子，分別進行栽培，並將們收穫後的豆子重量做成直方圖，結果也呈常態分配。如果豆子的大小是遺傳的，這二個曲線應該會呈二個平均值為中心的曲線，結果不是這樣，這二條曲線與它們祖先的曲線幾乎相同，看不出區別。

  (矮個子的父母生出高個子的子女，我們不是稱為「隔代遺傳」嗎？)

- 統計方法在了解自然本質上也有應用的好例子。一般人並不知道根據椰子樹樹葉螺旋的方向，可以分為右螺旋和左螺旋，這是遺傳特徵嗎？印度統計所的 T. A. Davis 進行了調查研究，他將不同螺旋形狀的樹木組合成雙親樹，並分類計算所產生的子孫樹具有相同特徵的數量。結果發現與基因遺傳無關，似乎是由隨機的外來因素決定的。不管雙親如何交配，子孫樹的右旋都略占優勢，約占 55%，是不是其生長環境中存在很大的可能性使得樹木的葉子向右螺旋？後來發現，從北半球收集的樣本中左旋率約為 0.515(即反時針的多)，南半球則為 0.473(即順時針的多)，恐怕是受到地球自轉方向的影響，與浴缸中放水的旋渦原理相同。也就是說，北半球的旋渦反時針方向的較多，南半球的旋渦是順時針方向的較多。

  Davis 還花了 12 多年比較一個大型種植園中左右螺旋樹的平均產量，驚奇的發現左旋的要比右旋的高約 10%。使他又聯想到在人類中，也有左撇子比較更有想像力、創造力、名人多等現象。更奇怪的是，除

了非常低級的以外，所有生物有機體的生化結構都是左手形的，在植物和動物的蛋白質中，甚至在簡單的有機體如細菌、病毒、霉菌中的氨基酸也是左旋形的，這些都值得更進一步從科學上去探討。

人類的大腦也分左腦和右腦，研究發現個體受右腦控制的人創造力較強，受左腦控制的人則更具邏輯推理能力，而受左腦控制的人是占多數的。

- 人的身高(真值)早上和晚上有差別嗎？如果有，差別有多大？要怎麼解釋？一項統計調查，分別在早上和晚上測量了學生的身高，發現早上的測量值要比晚上測的平均要多 9.6 mm，進一步研究發現這是因為白天脊椎骨之間的軟骨被壓縮造成的，夜間身體躺平時解壓後恢到原位，這稱為「日內循環」。事實上，人類還有其他的生理特徵也是在一天的 24 小時中有周期變化，每個人又有自己的內循環，這就是「時間生物學」，現已成為一個具有廣泛應用前景的研究領域。

- 1947 年印度剛獨立，德里發生了暴亂，大多數人避難到「紅色堡壘」的地方，少數人逃到休姆因廟裡。政府有責任提供食物給這些避難者，由於沒有正確的人數訊息，政府要如何核實承包商對休姆因廟各種生活保證品(例如豆、米和鹽)的帳單？統計專家利用承包商提出的豆、米、鹽的總量，再由消費調查平均每人每天所需的這三種物品的消費量，將總量除以單位消耗量，發現用鹽算出的人數最低，用米算出的人數最多，因為米價最高，所以人數就會被誇大，而鹽價非常低，因此不會誇大鹽的用量，從而得到了很好的人數近似值。

- 二戰期間，軍隊大量募兵時需要體檢，由於數量龐大，這是一項巨大的工作。假設其中有一項是驗血，如驗出某種疾病就必須淘汰。再假設平均 20 人中會有一個患此病，用什麼方法安排才能比一個一個檢驗節省驗血的費用？方法是：把 20 個人分為二組每組 10 人，然後檢驗同組 10 人的混合血樣，平均來說，會有一組呈陽性，然後對呈陽性的那組進行逐個檢驗，這樣的話，每 20 人的組平均僅 2 + 10 = 12 次檢驗，即減少了 20 次中的 8 次，少了 40 %。那麼如果把這 20 人再分成 5 個一組進行混合，則平均檢類的次數就只有 4 + 5 = 9 次，又節約了 11 次，即 55 %。當然，這種方法的最佳值要視該疾病的流行率而定。

  這個混合樣本的想法非常棒，可推廣用於其他領域，例如對自來水不同水源的水質檢測，確定是否有被污染。而且還能在不增加實驗設備的情況下，檢驗大量的樣本。現在，這種混合樣本的方法，已經被廣

泛應用於環保研究和其他領域中。