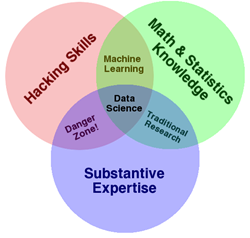
商研所碩一 鄭守開 (A09741303)

作業（一）

1. 將統計學(Statistics)稱為 Data Science。你是否同意?為什麼?

基於我過去的學習、工作、研究經驗，我不同意武斷地將統計學(Statistics)稱為資料科學(Data Science)，在資料科學這個新興領域當中，的確運用到許多統計學的知識、觀念、計算工具，但同時資料科學也包含了其他不在統計學範疇當中的其他領域知識，例如：電腦科學(Computer Science)的程式撰寫能力及對於特定領域的了解(Domain Knowledge)，因此我不認同將統計學(Statistics)稱為資料科學(Data Science)。



以上這張圖由美國數據科學家Drew Conway(Machine Learning for Hackers共同作者)製作，用於定義資料科學。由圖中可以看出資料科學為電腦科學相關技能(Hacking Skills)、數學及統計知識(Math & Statistics Knowledge)及大量專業領域的知識(Substantive Expertise)三者的交集，而數學及統計領域知識當中的一部分有被資料科學使用，因此我認為統計學中的一部分更像是資料科學的子集合，而並非可以用統計學這樣的字眼來取代資料科學名稱。

以下我也將提供我自身幾個學習、工作、研究經驗的例子證明我不認同將統計學(Statistics)稱為資料科學(Data Science)的觀點。

例1:

我自身目前也在計畫申請美國的資料科學研究所，因此花了相當多的時間研究不同大學對於統計學及資料科學的分界。我發現許多大學近年來都積極成立獨立的資料科學機構研究機構，如：紐約大學(New York University - Center for Data Science )及哥倫比亞大學(Columbia University in the City of New York - Data Science Institute)，與傳統的統計學系所獨立開來，也各自授予獨立的統計學(Statistics)、資料科學學位。而其他大學則是將資料科學放在傳統的統計系所，同時提供互相獨立的統計學及資料科學學位，如史丹佛大學(Stanford University)。因此我認為在美國的學術圈，目前也有逐漸將資料科學與統計學區分開的趨勢。

例2:

我在研究美國不同大學資料科學及統計學位時，也有查詢課程設計及必修課程，我發現資料科學與統計學的學位所要求的課程有所重疊，但並非完全相同。如：哥倫比亞大學的資料科學及統計學學位都要求修習機率(Probability)、統計(Statistics)及統計推論(Statistical Inference)等課程，但資料科學還額外要求學生必修更多電腦科學相關的課程，如：機器學習(Machine Learning)、演算法(Algorithms)等等，也有其他增加專業領域的選修課程，包括：資料視覺化(Data Visualization)、財金量化分析(Quantitative Finance)等等。因此我認為可以看出資科學領域除了統計知識外，也強調電腦科學相關技能及專業領域的知識等等。

例3:

我自身先前在中央研究院擔任研究助理職務時，時常需要做線性回歸(Linear Regression)、羅吉斯回歸(Logistic Regression)分析，當時我只認為這是單純的統計模型。但後來在學習資料科學、機器學習領域時，我發現除了電腦科學程式碼的部分之外，線性回歸、羅吉斯回歸也被大量的應用在機器學習領域，需多深度學習(Deeping Learning)模型背後都是使用這些統計模型作為工具。因此我認為資料科學這個新興領域當中，的確運用到許多統計學的知識、觀念、計算工具，但同時資料科學也包含了其他不在統計學範疇當中的其他領域知識，因此我認為統計學中的一部分像是資料科學的子集合。

1. 什麼是族群(population)? 什麼是樣本(sample)?

* 族群(又稱母體，Population): 為性質相同元素(Element)所觀測的全體資料，如：某牛奶工廠所產生鮮奶每瓶之容量、台北市30至40歲工作人口之收入。而當我們希望研究的目標族群稱作研究族群(Target or study population)。
* 樣本(又稱樣品，Sample): 為族群中一小部份的觀測值。當樣本結構及特性與族群相同時即為代表性樣本（Representative Sample）。而隨機樣本(Random Sample)則是族群中每個觀察值獨立且機會均等地被選取的樣本。

1. 什麼是參數(parameter)? 什麼是統計量(statistic)?

* 參數(又稱母數，Parameters)：為族群全部資料所計算出的數據，如：全台灣人口中女性的人數及比例、全台北市工作人口的平均收入。而族群參數通常以希臘字母代表之，如μ代表族群平均數。
* 統計值(statistics)：為樣本資料計算所得的數據，通常以英文字母代表之，如x̅上加一橫的X-bar，

代表樣品算數平均數。

1. 說說看如何估計台大全體學生每天平均上網時數。

理論上來說，若我們希望得到最為準確的台大學生每天平均上網時數，可以普查訪問每位台大學生進而得到精確的數字，但普查往往耗費大量人力資源及時間，因此實務上我們通常藉由隨機抽樣調查獲得一定數量的樣本，並在有科學根據的情況下，合理推測抽出的樣本沒有偏誤且具有代表性，可以有效地反映出整個母體(台大全體學生)的資訊(每天平均上網時數)。

法1:

根據教育部大專校院校務資訊公開平台，國立臺灣大學108學年度在學學生數為31,945人，若要逐一調查所有學生則相當耗費時間、人力成本，因此我們可以考慮隨機抽樣調查一部分的學生，如：31,945位學生的十分之一，約為3000人，藉由各種訪問方法，如：街頭隨機面訪、電話訪談、根據學號隨機發郵件訪談等等，通過訪問這部分的學生對母體(台大全體學生)做出合理的推估。

法2:

我們可以假設在學校上網的同學皆會使用NTU相關的wifi來使用網際網路，因此通過計算單位時間(一天)內的網路使用量/時數，再除以學校的總人數，進而獲得台大全體學生每天平均上網時數。但此方法有許多情形未考慮，而可能導致估計未能正確反映母體(台大全體學生)的資訊，如：台大學生未使用NTU相關的wifi來使用網際網路、學生可能同時使用不只一台電子設備連線使用wifi等等。