第一章 緒論

本章內容

1-1 統計學的歷史
1-2 統計學的應用
1-3 統計命題
1-4 母體與樣本

統計學(statistics)是一門數據科學，探討如何經由部分、有限的數據描繪整體的樣態，協助人們從事決策與判斷。廣義而言，每一個人依據手邊可得的資訊，或是個人累積的經驗，對事態發展判斷出最有可能的結果，這就是一種統計的應用。

我有幾次在週五下午約五點半的時候，搭計程車從新竹清華大學到火車站。其中兩次碰到了下雨天，第一次計程車塞在離火車站前約 500 公尺的地下道出口。走走停停花了好長一段時間，最後因為火車到站時間與計程車司機不願再更靠近車站的情況下，只好先行下車步行到車站。另一次，也是週五下班時刻，依然是驟雨剛過。這一次，這位計程車司機聽我要去車站，就警告會塞車。有了上次的經驗，我完全贊成這位司機的說法，我當下決定不去前站，我們往後站人行地下道入口下車。

每個人每天都在不斷累積各種經驗，就像一個超大的硬碟被不斷地儲存各種資料。但這些資料是死的，必須要有一些特有的方法將這些資料整理、分類和轉化，方能形成屬於你或者我可以運用的知識。這些因人而異的特有方法，造成了彼此決策上的差異。所以，會有司機塞在大雨淋漓的街道上，有些司機卻也能因為下雨天，乘客變多而多賺一些。

統計學提供處理數據的科學方法，以輔助人們更為精確地進行決策，精確的意涵不意謂是完全正確或是最佳的的決定，而是指非常接近只有上帝才知，或未來才知的正確答案。

1-1 統計學的歷史

統計一詞的英文原文為 statistics，源自拉丁文 status，意為狀態。乃指一個國家的各種政經和社會狀態，內容廣泛包含經濟，人口，牲畜，農穫，軍事等等。現今政府蒐集的資料更是多樣，這些包含進出口貿易，死亡原因的分類，氣象資訊等等。然而，如果沒有一些有效檢視這些資料的方法，這些也僅只是資料的蒐集罷了，不能提供決策的參考。

當統計不再只是停留蒐集和紀錄各項資訊，而是嘗試從這些資料去描述一個整體型態，才逐漸形成一門獨立學門。我們很難說統計學的發端起於何時，可以說最近一百年，也可追溯到更早的時期。統計學這套方法論有別其他思維的最大特色，就是體認到 “誤差”的存在。法國數學家拉普拉斯(Pierre-Simon de Laplace, 1749-1827)也是一個天文學家，他在利用觀察數據預測行星與彗星的位置時，引入 “誤差函數”(error function)的概念來解釋預測值與實際值有偏差的問題。當時，對於天體位置的預測誤差歸因於觀察誤差，普遍認為如果觀察的機具更為進步，就可以克服估計誤差的問題。當代統計學，除了認知道量測或觀察誤差的存在， 在很多領域的觀察，發現數據具有“不確性”(uncertainty)。例如，我們量測從家裡到學校的距離，這個距離是固定不變的，但每次測量值卻會發生變化，這大部份歸因於量測誤差。然而，若是記錄每天從家裡到學校的時間，每天所花的時間不盡相同，而這樣的差異不是量測造成的，而是各種因素影響下所構成的“不確性”。

1-2 統計學的應用

統計學的核心價值在於可以整體對象作出一般性的描述，也因而可以發覺異於整體型態的個體。這樣的概念可以在很多領域都有成功的應用經驗。

農業改良與疾病預防：

統計學對現今最重要的貢獻就是對提高農產品產量，使得在有限的土地資源下，可以得到最多食物。早期的農業實驗探討很多因素對產量的影響，這些包含肥料與施肥方式，灌溉，病蟲害防治以及品種選擇方面等。

工業統計：

1924 年舒華特（Walter A. Shewhart）根據統計學檢定（statistical testing）的概念提出統計製程管制（statistical process control, SPC），用以監控製程是否出異常。一個製程如果處於異常狀況將使產品的品質下降，因此，如果可以避免異常狀況持續，就可以提升產品品質，減少報廢與降低成本。另外，早期農業實驗所發展出來的方法，也可以應用在工廠上，讓產品設計開發階段就可以瞭解各種影響產品品質的因素。例如，一個化學製程，需要管控生產流程的壓力，氣體濃度，流速等因素使得生成的產品品質最高，並瞭解一些可控變數對產品品質的影響。

醫學統計：

在流行病學（Epidemiology)上，常用回溯性方法蒐集資料，探討一種行為與否、區域性差異與疾病的關係。例如，現代醫學已經證實抽菸與肺癌存有關連性，抽菸的人得到肺癌的機會遠高於沒有抽菸的人，吃檳榔與口腔癌也有關聯。除此之外，醫療檢驗也用到統計陳述其檢驗敏感性[1]與特異性[2]。又如，新流感 H1N1 的傳播方式為接觸與飛沫傳染，檢驗出陽性的機率約為 50%~60%等。

1-3 統計命題

統計學還可以應用在很多領域上，端賴各位如何去描述你所處的問題型態與轉換成一個統計可以回答的問題。例如，統計學可不可應用在文章作者的辨識上？假如有一篇文章，作者可能來是甲，也可能是乙，如何辨識呢？當然，專業人士可能從某些特徵是否存在加以論證，嚴格來說，這些近似先驗的知識也是一種統計的結果。所謂特徵，它可以字句的長短分配，標點符號的使用，特殊習慣等。這裡所謂的特徵就是選擇了一個角度去說明何謂“像”，何謂“不像”也就是選擇了一個角度去說明兩者差異的距離。又如，大學入學考試的科目就是一種角度的選取，不同學校相同科系甚至採計不同科目計算錄取分數。這反映出我們對於評估學生能力的方式有所不同，這些差異當然影響學生是否被錄取。

[1]敏感性(sensitivity)：疾病發現之能力，有病者被驗出陽性的比例。

[2]特異性(specificity)：無病發現之能力，沒有病被驗出陰性的比例。

然而，選擇何種角度並不是統計問題，卻比統計本身來得重要，統計學只是在給定的 “角度”下去辨識差異是否存在，差異是否明顯等問題。所以，在運用統計分析的結果，必須很小心，有限的資料隱藏者它的限制，不可做過分的推論。

1-4 母體與樣本

一個生產信用卡卡片的公司，想要瞭解該公司信用卡的壽命。壽命一詞常只一個產品從出生到毀壞的時間長度，但這裡壽命通常是只它可以被刷幾次而言。在這個問題上，明顯地陳述整個未來資料蒐集的範圍，就是該公司的信用卡。一個研究首要確認研究對象的範圍，這個範圍內的成員構成這個研究問題的母體(population)。確認母體就是為研究範圍畫了一個界線，也為研究結果的推論畫了一個界線。所以，該公司產品的壽命不能推論到其他公司信用卡的壽命。有如，一個市場調查想要探究桃園縣中學生每日就寢時間，這個市場調查的母體就是桃園縣中學生。簡言之，**母體就是所有研究對象所成的集合。**

當然，我們不可能去測試該公司所有信用卡的壽命，我們也不能調查所有中學生的就寢時間，這樣的工作不僅無法符合實際需要也費力耗時。所以，我們必須選擇一部分的產品做測試，調查一部分的中學生。這些被測試的產品，受調查的學生我們稱為樣本（sample)。簡言之，**樣本就是母體的部分集合。**

每一張信用卡的壽命都不一樣，即便它們都是來自相同的公司與生產方式，我們也盡可能將所有生產條件控制一樣，但還是無法避免信用卡的壽命存有差異，稱此為變異（variation)。一個母體內的每一個成員都不一樣，如前所述，我們用分配來表現這些差異情形。但一種更為簡略的方式，就是用一些有意義量化方式來表現這個母體的特點。我們稱此描述母體特性的測量值為參數(parameter)。常用的參數有母體平均值（mean)，中位數（median）等等。我們可以比較兩家信用卡公司產品的平均壽命，決定採用哪一家的產品。

然而，因為我們沒有調查所有母體成員，就不能確實知道母體參數之值，所以，參數通常為一個未知數。而統計學就是想要利用樣本資訊去估計這個未知參數的值，而由樣本所計算得出的數值，稱為統計量（statistic)。一個用於估計某一個參數的統計量，稱為該參數的估計量（estimator)。

母體(population)：所有研究對象所成的集合。
樣本(sample)：母體的部分集合。
參數（parameter)：描述母體特性的測量值為。
統計量（statistic)：由樣本所計算得出的數值。

例 1- 1

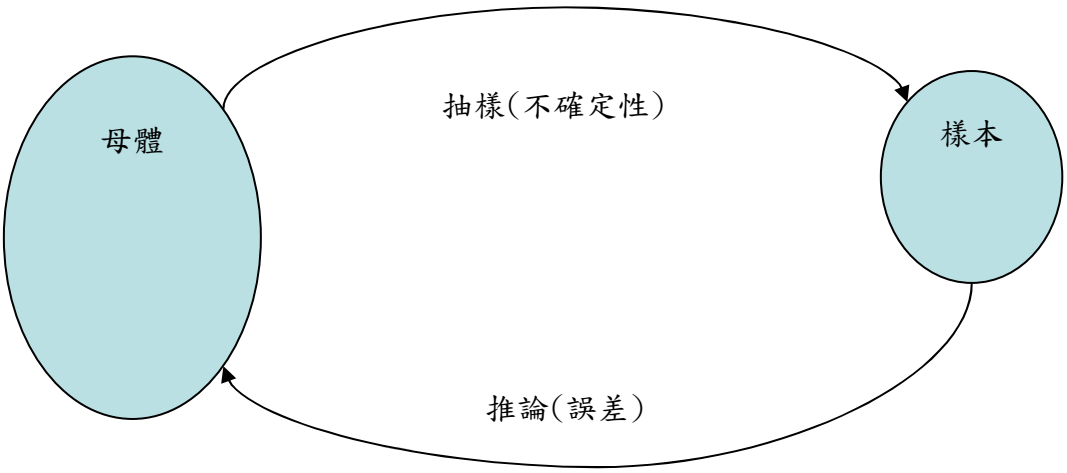
銘通信用卡製造公司欲調查該公司產品的壽命，他們隨機抽出 10 張信用卡，測試它們連續刷卡的次數，測試結果為：

1019　995 1027 1019 1010　987 1021　986　994 1000

該公司產品的平均壽命為一個未知參數。該公司品管人員建議採用樣本平均值來估計，樣本平均值為

$$\frac{1019+995+1027+1019+1010+987+1021+986+994+1000}{10}=1005.8$$

假設母體真正的平均值 1000, 估計值與真值間有差距，稱為估計誤差。估計誤差來自於我們選取信用卡的抽樣過程，抽樣具有不確定性，才使得估計存有誤差(error)。所謂抽樣的不確定是指 “若抽樣行為可以重複進行，每一次抽出的信用皆不相同”不相同的信用卡測得的壽命資料就不一樣，所計算的平均值也就可能不同。實際上，真正的參數值是未知的，所以，每一次的估計誤差為多少也是一個未知數。不過，我們可以利用機率的抽樣分配理論，評估誤差的可能範圍，就可以知道用樣本資訊推論母體參數是否可信。



定義：

估計誤差=估計值-真值。

參考資料

主計處 http://www.dgbas.gov.tw/ct.asp?xItem=13213&CtNode=3504

衛生署疾管局 http://www.cdc.gov.tw/mp1.htm

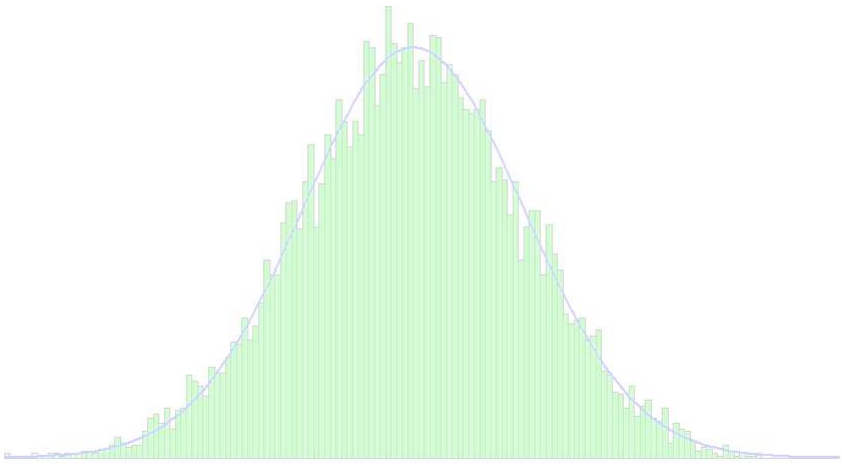
自由時報 www.libertytimes.com

TVBS www.tvbs.com.tw

習　題

1. 何謂母體？何謂樣本?舉一例說明其間差異？
2. 何謂參數？何謂統計量？請舉一例說明它們的關係？
3. 針對『澎湖博弈公投』的選民意向，TVBS 於 2009/9/18,21,22 三天晚上，針對澎湖 20 歲以上的公民進行電話隨機民調，有效樣本 956 位，在 95%信心水準下，正負誤差為 3.2%。受訪者中，有 49%反對觀光賭場，只有 38%贊成。請問上述問卷調查的母體，樣本，參數和樣本統計量為何？根據調查結果，你有何看法？

4. 根據自由時報 2009/9/27 報導：『台灣有史第一次的博弈公投昨天舉行，選前外界評估正反雙方五五波，結果不同意票大勝近四千票、以五十六．四四％比四十三．五六％的比例過關，跌破專家眼鏡。公投沒過關的關鍵之一是公教軍警聚集的馬公市，不同意票大贏三千八百票。這場博弈公投，正反兩方角力不斷，在沒有投票人數門檻、贏一票就算贏的情況下，雙方動員積極。澎湖縣公民數七萬三千六百五十一人，投票率四十二．一六％，高於原先預估的三成五至四成。投票結果為不同意票數一萬七千三百五十九票，同意票一萬三千三百九十七票。』請分析第 3 題 TVBS 民調的估計誤差。



重點摘錄