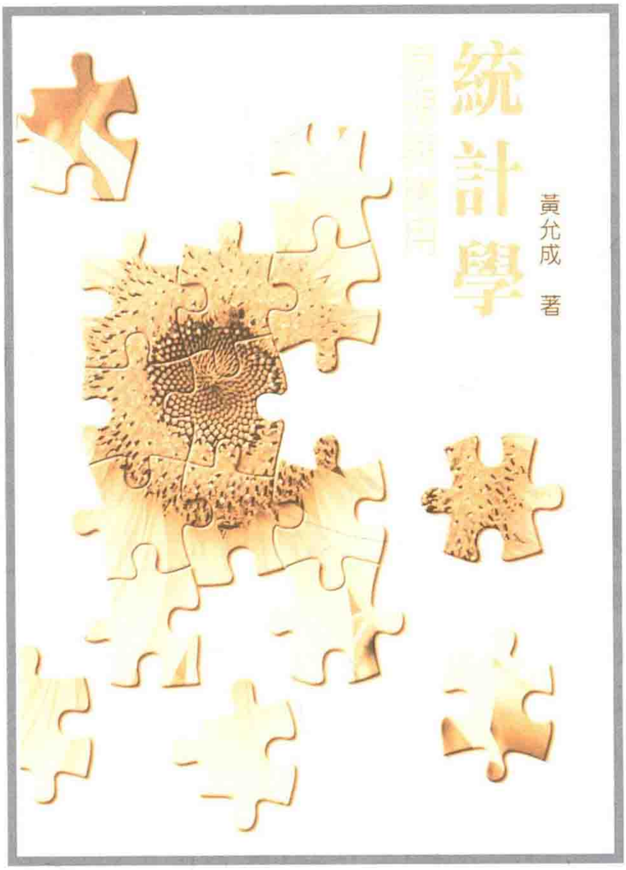
統計學

原理與應用

黃允成 著



統計學

原理與應用

黃允成 著

雙葉書廊
YEH YEH BOOK GALLERY

國家圖書館出版品預行編目資料

統計學：原理與應用 / 黃允成著. --
初版. -- 臺北市：雙葉書廊, 2013.06
面； 公分
ISBN 978-986-6018-56-5（平裝）

1.統計學

510 102011570

統計學：原理與應用

作　　者　黃允成
發 行 人　張福隆
責任編輯　魏安邑
封面設計　呂秀蓉
出 版 社　雙葉書廊有限公司
地　　址　台北市羅斯福路三段 269 巷 12 號 1 樓
電　　話　(02)2368-4198
傳　　真　(02)2365-7990
網　　頁　http://www.yehyeh.com.tw
讀者服務　pub@yehyeh.com.tw
登 記 證　局版北市業字第 239 號
出版日期　西元 2013 年 6 月 初版一刷
電腦排版　佳軒電腦打字社

ISBN：978-986-6018-56-5

自序

個人從事教學生涯已逾 20 個年頭，從教學的第 1 年開始就教「統計學」，換言之，我教「統計學」也已超過 20 年，雖說坊間的統計學教科書大多都有一定的水準，但仍覺未能盡如人意，例如有的太簡單，對理論背景沒有完整交代；有的太繁雜，雖對理論背景有所交代，但章節內容過多，若要一一學習，恐要 2 年才教得完；有的又太難，不太適合初學者。基於此，本人決定著手寫一本適合初學者，又能將理論背景做一完整交代，且對各章節內容都能賦予適當數值範例，以利理解的統計學教科書。自我從事教學以來，我就不太喜歡採用翻譯的教科書，理由是因為國情不同、環境不同，且翻譯的品質又良窳不齊，有些範例又不見得適合於國內情況，因此，除非該原文教科書寫得非常好而直接採用外，實在不太適合拿翻譯書籍來當教科書。然而，採用原文教科書也有它的缺點，那就是「文字障」，許多學生英文程度不盡理想，閱讀原文教科書相當吃力，可能會造成他的學習障礙，也許他原來對統計學是有興趣的，但因語文障礙而連帶地影響他對統計學的學習熱誠，那就得不償失了。在教學過程中我常跟學生提及，本人的一生夢想是渴望有一天，我的名字能被編入統計學教科書之中（例如柴比雪夫（Chebyshev）、馬可夫（Markov）、皮爾森（Pearson）、卜瓦松（Poisson）、貝氏（Bayes）、……等等），但合理預期，此一夢想可能要落空了。因此，退而求其次，寫一本統計學教科書，就成為本人唯一可以聊表自慰的替代選擇了。本書將各章節所用到的公式或定理，盡其可能地加以證明或推導，其目的是希望學習者能「知其然，亦能知其所以然」，不要只將公式或定理，視為背誦的標的而不去理解，另一方面，在各章節中加入適當的數值範例，以利於相關章節內容的理解與應用。本書歷經 1 年多的撰寫，雖力求嚴謹，但因個人能力有限，疏漏在所難免，尚祈各界不吝賜教與匡正。

黃允成

目錄

chapter 1

統計學導論

本章大綱

1.1 導論：統計學是什麼？統計學到底在學什麼？

統計學是「統而計之」之學。「統而計之」是什麼意思？「統」是指有一堆資料，將它們統合在一起。「計」是指將它們拿來加以計算，以取得一些具有意義的資訊。其目的為何？就是希望經由這些有意義的資訊，推論這一堆資料背後潛藏的真相。所以統計學是希望利用所取得的資料，透過數學計算與分析的方法，推論這堆資料其所來自的母群體的可能狀況。在此，我們定義「母群體的可能狀況」稱之為「真相」。實際上，所謂的「母體真相」是指「母體的某一個統計特徵量數」而言。「母體真相」又稱為「母體母數」或「母體參數」，例如母體平均數、母體百分比或母體變異數等等，都是母體母數，或稱為母體參數。所以統計學又稱為「真相之學」。

統計學主要分為兩大類，一為「敘述統計學」，一為「推論統計學」。「敘述統計學」是指將所取得的資料加以計算、彙整，用以描述該資料的現有狀況，而不去推論該資料背後的母群體真相。「推論統計學」則是利用所取得的資料，經由計算與分析，用以推論該資料背後所潛藏的母群體真相。所以在推論統計學中，我們將所取得的資料稱之為「樣本」，將所欲推論的母群體稱之為「母體」。

所謂「母體」是指主事者有興趣研究的主題所含蓋的全體。母體大小一般是以 N 來表示。所謂「樣本」是指母體的部分集合。樣本大小一般是以 n 來表示。例如如果你有興趣研究的是中山大學學生購買手機的消費行為特性，那麼，你的母體就是中山大學的學生，而如果你從中山大學學生中隨機抽取 300 人進行問卷調查，則這 300 人就是你的樣本。如果你有興趣研究的是全國大專院校學生購買手機的消費行為特性，則全國大專院校學生就是你的母體，而如果你從全國大專院校中隨機抽出 10 所大專院校，再從這 10 所大專院校中再隨機抽取 1200 人進行問卷調查，則這 1200 人就是你的樣本。由此可見，母體並沒有一定的大小，而是視研究主題所含蓋的範圍而定，母體一經確定，則樣本就可依此加以明確定義，即樣本是母體的部分集合。

在進行統計推論時，主要是經由樣本資訊去對母體真相進行預測。因此，我們希望樣本必須具有代表性，否則，依此所做的推論必然會產生偏差而無法做出有效的推論。所謂樣本的代表性又是什麼意思呢？簡單地說，如果樣本是母體的具體而微，那麼，該樣本就具有代表性。那麼，具體而微又是什麼呢？具體而微係指樣本是母體的縮小版，就如同幾何圖形的相似性一樣，母體是較大的圖形，而樣本是較小的圖形，兩者的圖形相似，只是大小不同而已。什麼樣的樣本才具有代表性呢？要使樣本具有代表性，就必須對母體的結構與特徵有一定的認識。對母體的結構與特徵有一定的認識之後，才能選擇正確的抽樣方法，經由正確的抽樣方法才能產生具有代表性的樣本。有了具代表性的樣本之後，再經由適當的統計推論方法與程序，才能對母體的真相做出較為準確的判斷。茲將統計學的整體概念架構圖圖示於後：

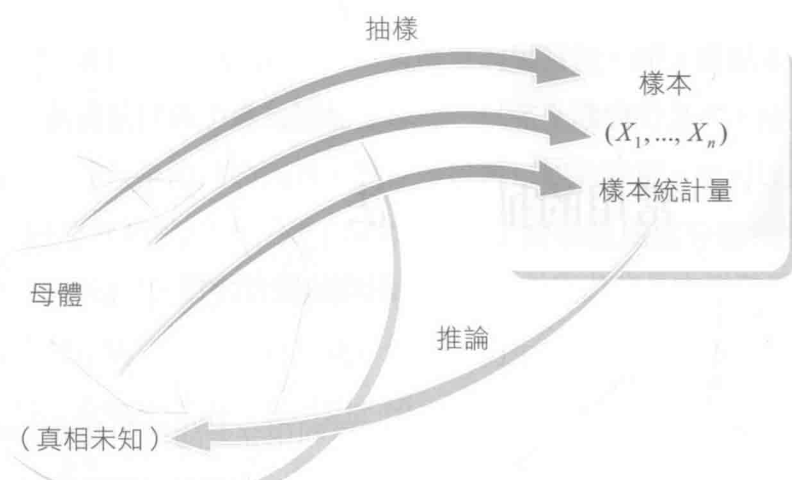


圖 1-1 統計學整體概念架構圖

從圖 1-1 可知，統計學的主體概念是想要經由抽樣，去推論未知的母體真相。換言之，如果母體真相已知，那麼，就不必用到統計學。上圖中有一未經定義的專有名詞「樣本統計量」，所謂「樣本統計量」係指樣本資料的函數關

係，以數學符號表示為：

$$\hat{\theta} = f(x_1, x_2, ..., x_n)$$

其中：$\hat{\theta}$代表樣本統計量，x_i表示第 i 個樣本的資料值。

利用樣本統計量對母體真相進行推論時，可能會做出正確推論，也可能做出錯誤推論。既然如此，那與一般的猜測何異？一般猜測的結果也是可能正判，也有可能誤判。其間最主要的差異在於「利用統計推論可以事前知道，在既定的誤差水準之下，正確地推論母體真相的機率有多少」，亦即統計推論不是一般的隨意猜測，它能告訴你在可容許的誤差範圍之內，正判的機率有多少。如果主事者對此正判的機率並不滿意，那麼，可以再利用統計方法，計算出能滿足所需機率下的樣本個數（Sample size），重新計算樣本統計量，再進行必要的統計推論。

1.2 常用的抽樣方法

為了讓樣本具有代表性，我們常用的抽樣方法有：

一、簡單隨機抽樣法

此法是將母體視為一單一的母體，再利用完全隨機的抽樣方法，抽取所需的樣本數，以作為後續統計推論之用。此法一般是利用亂數表或電腦化的亂數產生器，產生所需的亂數，作為抽樣之用。採用此法時須將母體中的各個個體加以編號，或經由系統化的方法將之加以組織，以方便後續抽樣之用。例如母體大小為 1000 個個體，今欲從其中隨機抽取 100 個樣本，那麼，就須將母體中的每一個個體加以編號（1~1000 號），再利用亂數產生器，產生 100 個隨機亂數，該亂數的值恒介於 1 到 1000 之間。

二、分層比例抽樣法

此法是假設母體可依某種分類方法，將其分為數個次母體，再依次母體佔總母體的比例，抽出各次母體所需的樣本數，作為後續統計推論之用。依法所得之次母體稱為分層次母體，層內之次母體具較高之同質性，不同層次之間之次母體則可以有較大的異質性，即「層內差異小，層間差異大」。例如母體依性別可分為兩個次母體，若男女兩個次母體佔總母體的比例分別為 7/10 和 3/10，且若總樣本數為 1000 個，則男的樣本數應為 700 個，而女的樣本數則為 300 個。

三、叢式抽樣法

此法是用在母體中具有多個相似集群的次母體的情況下，亦即母體是由多個類似的集群（次母體）組合而成，該次母體稱之為「叢」（Cluster）。換言之，母體本身是一個大叢，它是由數個小叢所組成。每一個小叢之間具有高度的結構相似性，但叢內的各個個體則可以有高度的異質性。即「叢間差異小，叢內差異大」。此法是以各小叢為被抽樣單位，以完全隨機的方式，抽出任意一小叢作為樣本，進行全面性的調查，以推論母體可能的真相，也可以再從所抽出的小叢，根據其他抽樣方法進行抽樣（例如簡單隨機或分層比例抽樣等）。例如若欲調查高雄市市區國中學生的補習情況，則由於高雄市市區之國中學校具有高度的結構同質性，有三個年級，有男生有女生，有常態編班之班級，也有各分科之資優加強班，有用功的學生，也有不用功的學生，有乖的學生，也有不乖的學生，有家境好的學生，也有家境不好的學生，……等等。換言之，不同的學校之間具有結構上之同質性，而學校內部則具有高度的異質性。因此，在抽樣上可以學校為被抽樣單位，隨機抽出一個學校作為樣本，再對該學校進行全面性的普查或抽樣調查，以作為母體推論的依據。

四、系統抽樣法

此法是將母體中的各個個體進行編號（1 到 N），然後決定出所需之樣本

數（n），再將母體分割成 n 個次母體，每一個次母體內包含有 N/n 個個體，今從亂數表或電腦亂數產生器中產生一個介於 1 到 N/n 的隨機亂數（y），作為第一個次母體的抽樣樣本，而第二個樣本則從第二個次母體中的同一位置抽出，即編號為（$y+\frac{N}{n}$）的樣本，第三個樣本則從第三個次母體中的同一位置抽出，即編號為（$y+2\cdot\frac{N}{n}$）的樣本，依此類推，可將 n 個樣本全數抽出。例如高雄市市民對張三競選高雄市市長的民意支持度之電話調查，就可以利用電話簿進行系統抽樣，方法是先根據允許的抽樣誤差算出所需的樣本數 n，再將電話簿分成 n 個等分，然後，用亂數表或電腦亂數產生器產生一隨機亂數，作為每一等分在同一位置之取樣樣本，再據以進行電話調查，取得所需的樣本資料。

取得所需樣本之後，就須將其拿來加以計算，以產生我們所需要的樣本統計量，然後，再根據樣本統計量去推論母體可能的真相。統計推論的方法，大體而言分為兩大類，一類稱為「參數估計」，另一類稱為「假設檢定」。「參數估計」又可分為「點估計」及「區間估計」。「點估計」是指以樣本統計量直接去估計母體的真相，屬於單點估計的方法，而「區間估計」是以樣本統計量為估計的基準點，再加減一個抽樣誤差的範圍，形成一個區間，然後，以該區間作為估計該母體真相的可能範圍，即母體真相在一定的誤差範圍內，有一定的機率（如 95%）會落在該區間的範圍內。截至目前為止，我們並沒有對母體真相有過更詳細的探討，實際上，母體真相是指母體的某些統計特徵或重要的屬性狀態而言，這些母體的特徵及重要的屬性狀態一般又稱為母體參數（或母體母數）（population parameters），這些母體參數是主事者所感興趣或對其具有重要意義的屬性狀態。例如張三想要參選台北市長，那麼，張三就是主事者，他感興趣的是他是否會當選？然而，要如何才能得知張三是否會當選呢？最理想的方法是透過民意調查，以了解其民意支持度高低，以作為是否會當選的判斷依據。因此，就民意調查而言，台北市市民就是母體，而台北市市民的民意支持度就是母體的真相，也就是母體參數。

一般而言，「點估計」的方法是較難準確命中母體真相的，這就像是打靶一樣，要一顆子彈剛好命中紅心其實是比較困難的，除非槍跟槍手都是非常優秀的，否則實屬不易。「區間估計」則改以區間的方式來估計母體可能的真相，其命中的機率相對較高，但區間估計法也潛藏一個問題，那就是「只要區間加大，則母體真相落在區間裡面的機率自然增加，甚至可以到達百分之百的地步」，然而，區間加大背後代表精確性下降，模糊性增加，對主事者而言，以該區間作為輔助決策的依據，其效果將大幅地下降。也就是說，以擴大估計區間作為提高命中率的手段並非主事者所樂見。主事者期望的是「估計的區間小一點，但命中母體真相的機率要高一點」，要達到這個期望目標就必須增加樣本數，而增加樣本數就會增加抽樣成本。換言之，要在既定的命中機率下，縮小估計區間，就必須以增加成本支出為代價，這是「魚與熊掌不可兼得」的問題，主事者必須權衡利弊得失，在成本效益的綜合考量下，計算出最適的樣本大小，以作為統計推論的依據。

至於「假設檢定」，它是將母體真相的所有可能的狀況，分割成兩個「周延」而「互斥」的假設。所謂「周延」是指這兩個假設的聯集就是母體真相的所有可能的狀況。所謂「互斥」是指這兩個假設的交集為空集合。假設檢定的兩個假設，一個稱為虛無假設（H_0），一個稱為對立假設（H_1），根據上述的定義，則 $H_0 \cup H_1 = \Omega$ 且 $H_0 \cap H_1 = \varphi$，其中 Ω 代表宇集，φ 代表空集合。所謂「宇集」是指針對該特定問題或隨機實驗的所有可能出象所成的集合。所謂「空集合」是指該集合中無任何元素之謂也。假設檢定就是根據樣本資訊去判斷母體真相到底是落在 H_0 或 H_1。一般而言，H_0 與 H_1 是以二分法進行命題的，兩者不但沒有交集，而且母體真相不可能同時落在 H_0 與 H_1 裡面，也不可能同時不落在 H_0 與 H_1 裡面，也就是說母體真相必然落在兩個假設中的其中一個。例如法官想要判斷張三是不是台東搞軌案的凶手，他須建立兩個互斥且周延的假設：

H_0：張三不是搞軌案的凶手

H_1：張三是搞軌案的凶手

這兩個假設互斥且周延，真相只會在其中的一個假設內，不會兩個假設同時成立，也不會兩個假設同時不成立。假設檢定方法是在建立假設之後，必須蒐集更多資料，然後加以整理分析，最後判斷兩個假設中何者成立，何者不成立。即到底是 H_0 為真，還是 H_1 為真。做完判斷之後，該判斷結果可能是正確的，也可能是錯誤的。正確判斷是我們所期望的，而錯誤判斷則是我們所不樂見的，既然錯誤判斷是我們所不樂見的，那麼，我們就必須設法降低錯誤判斷發生的機率，才能達到我們預期的目標。在假設檢定上，錯誤判斷又分為兩種類型，一種稱為型一誤差，另一種稱為型二誤差。所謂型一誤差係指在 H_0 為真的條件下，主事者卻拒絕 H_0 而接受 H_1。即真相為 H_0 為真，但主事者卻判斷 H_1 為真。所謂型二誤差係指在 H_1 為真的條件下，主事者卻拒絕 H_1 而接受 H_0。即真相為 H_1 為真，但主事者卻判斷 H_0 為真。茲以表 1-1 說明此一情況。

在上述例子中，假設張三的確是搞軌案的凶手，但因無法蒐集到足夠的證據以定他的罪，所以將其無罪開釋，這種誤判的型式稱為型一誤差；反之，如果張三的確不是搞軌案的凶手，但因遭人陷害，種種證據都不利於他，而被認定有罪，這種誤判的型式稱為型二誤差。假設檢定的方法之所以成為一門科學的方法，是因為它不但須要大膽假設，更需要小心求證，使其在既定的誤差範圍內，做出正確的判斷。也許有些人會說，為什麼不能做出百分之百的正確判斷呢？這是因為在科學的論證上，除非有百分之百的證據可以確定母體的真相為何，否則，都不能武斷地說，我的判斷是百分之百地正確。就統計推論而言，除非進行母體普查而非只是抽樣調查，否則，都無法百分之百確信我們所做的判斷完全正確。然而，母體普查在很多情況下是不可能實現的，例如破壞性品質檢驗，由於該檢驗具有破壞性，產品一經檢驗即告損壞而無法販售，因此，不可能採取百分之百檢驗，據以確定其整體的品質狀況。又例如法官之辦