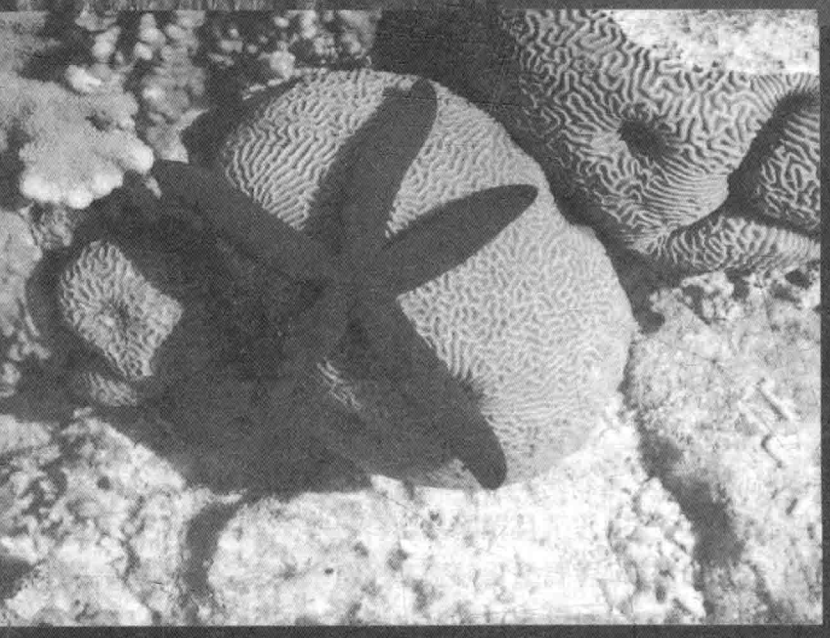
生物統計 第二版

AN INTRODUCTION TO BIOSTATISTICS Second Edition

Thomas Glover · Kevin Mitchell 原著
方嘉德 譯

滄海圖書
Tsang Hai Book Publishing Co.



生物統計 第二版

AN INTRODUCTION TO BIOSTATISTICS Second Edition

Thomas Glover · Kevin Mitchell 原著
方嘉德 譯

滄海圖書
Tsang Hai Book Publishing Co.

國家圖書館出版品預行編目資料

生物統計 / Thomas Glover, Kevin Mitchell 原著；方嘉德譯. -- 二版. -- 新北市 ： 滄海圖書資訊, 民 102.03
面； 公分
譯自：An introduction to biostatistics, 2nd ed.
ISBN 978-957-28907-8-3 (平裝)

1. 生物統計學

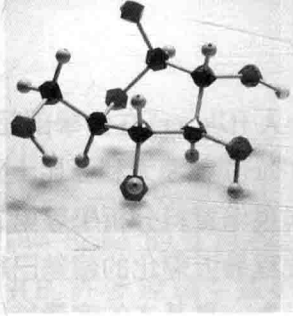
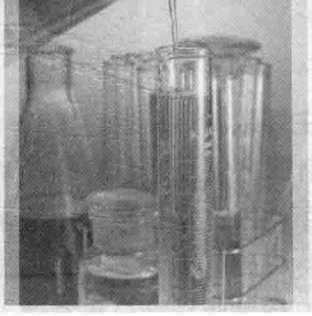
360.13 102003693

滄海書碼 PS0356C

生物統計 An Introduction to Biostatistics 2e

原著者 ／ Thomas Glover, Kevin Mitchell
譯 者 ／ 方嘉德
發行人 ／ 陳展維
出版者 ／ 滄海圖書資訊股份有限公司
總經銷 ／ 滄海圖書資訊股份有限公司
地 址：23552 新北市中和區中正路 679 號 6 樓
電 話：(04) 2226-3103
傳 真：(04) 2226-3156
網 址：http://www.tsanghai.com.tw
E-mail：thbook@tsanghai.com.tw
中華民國 102 年 5 月二版一刷

ISBN 978-957-28907-8-3

作者序

PREFACE

本書撰寫的目標是，希望製作出對於需要在生物科學領域使用到統計學的大學部學生而言，這將是一本具有容易理解且相對完整的入門介紹性書籍。本書也是針對生活科學領域的一季或一學期之基礎統計學課程所設計。預設的目標學生是生物學、環境研究科學、生物化學和主修衛生科學等各領域的二、三年級學生。我們假設研讀本書應具有的背景知識是修完一些生物學和基礎代數學等課程，但是不一定要修過微積分。在本書中，我們列舉出了包括遺傳學、生理學、生態學、農業和醫藥學等各種生活科學領域的一些範例。

本書強調在機率、機率分布和假設檢定等各項之間的關係性，同時也強調在虛無假說和研究假設等各項中的各種檢定統計量之期望值，以做為理解假說檢定方法論的一種方法。此外針對許多情況，除了標準的參數分析方法之外，也導入無母數型的替代方法。會涵蓋這些無母數技術的原因是，大學部學生的學習報告通常只有較小的樣本大小，因而會先排除掉參數分析方法，而且對於具有適當數學背景的學生而言，無母數檢定方法的推展過程是相當容易被理解的，故忽略或跳過無母數方法的作法，並不會影響到本書內容的連貫性。

在本書每一種觀念之介紹文章中，都嘗試舉出一些有趣而容易理解的範例。而在每一章的結尾部分，各項問題涵蓋了不同的難易程度，並且也是取自於各種知識領域。其中大多數並不是實際的生活例子，但是題目設計和數據數值等各項則都是具有寫實性。在每一章結尾部分的各項問題是採用隨機安排的方式，因此要求學生需選擇出適當的分析方法。許多大學部教科書則是先介紹一種觀念或檢定方法之後，就立刻提供出許多問題，而全都需要使用這種技術來解答，但這種作法阻礙了學生在現實生活中做出選擇適當分析方法決策的學習過程。筆者們相信，在統計分析方法的介紹中，學習做出這種判斷決策乃是一種關鍵技巧，因此本書中提供了很多機會，讓學生能夠練習而且發展出這種技能。

這本書的內容主要取材自筆者的其中一人 (Glover) 在大學部超過 20 年期間所教授的生物統計學，以及另一筆者 (Mitchell) 在澳大利亞昆士蘭之海外課程的最近一些期間所教授之無母數統計學與現場資料分析學等第二種課程的相關內容。對於生物學學生而言，大學部課程最近變化的趨勢已經不再強調微積分，反而目前是強調將統計分析方法做為一種基本的定量技能。因此，筆者們希望這本教科書能夠讓教導與學習這項技能的過程變得不再是那麼地艱難。

補充教材

有許多統計套裝軟體和輔助教材能夠支援本教科書中的教授內容。選擇這些支持教材與否，端視個人的興趣和成本考量所決定。目前在筆者們之教學實作課程中使用了 SPSS 套裝軟體 (http://www.spss.com/)。這套軟體是相當容易上手，且是相對較有變化彈性的，而且幾乎可以操作在本教科書中所介紹的所有統計技術。同樣有用且容易到手的其他統計套裝軟體程式為 Minitab (www.minitab.com)。也有許多免費的線上統計工具可供利用。在這些最好與最完整之軟體中，其中一種是可在 http://www.r-project.org/ 網站下載的 The R Project for Statistical Computing 軟體。

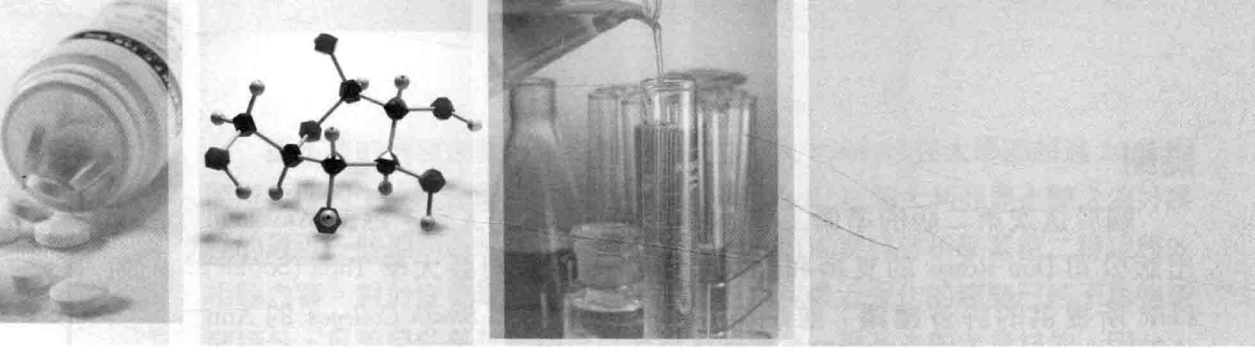
對於學生而言，我們會建議購買德州儀器公司所生產而型號從 TI-30 到 TI-83，TI-84 和 TI-89 的各型計算機。這些計算機的定價差距很大，而且可能成為在特定課程中選擇一台必需計算機時所需要考量的一項因素。雖然像 TI-30 的計算機很難自動化運算，但是藉由需要使用更多的計算步驟，有時候它們能夠讓學生對統計檢定方法有更清晰的洞察理解。當電腦程式或精密電子計算機更容易執行計算程序時，有時候會對統計學和其相關的計算過程，在腦力思考上產生了一個「作業黑箱」。

對於學生和教師們而言，筆者們推薦以下一本書籍：D. J. Hand et al., editors, 1994, *A Handbook of Small Data Sets*, Chapman & Hall, London。這本書包含了在 1875–1894 年之間被馬匹踢死的普魯士軍人 (data set #283) 到 Shoshoni 印第安人之皮革貨物上的珠串形狀 (data set #150) 等 510 個小型資料組。這些資料組很有趣且很容易處理，同時也可以使用本教科書所介紹的技術來進行統計分析處理。在本書中，已經使用這本手冊中的許多資料組來做為內文範例或章節的問題，此外也還有許多的其他資料組可以被轉變為迷人而有用的實務問題。

誌謝

關於這次第二版的準備工作，我們要感謝下列人士：感謝 Waveland 出版公司 Don Rosso 的支持和指導；感謝南澳大利亞大學 Topa (Sophie) Petit 所提供的許多建議；並感謝 *Hobart and William Smith Colleges* 的 Ann Warner 之字斟句酌，而讓這份手稿有了各式不同的輪廓；此外，也感謝 Hobart and William Smith Colleges 的學生所提出的許多意見和建議，特別是 Aline Gadue 對第一版所做的仔細校閱。

Thomas J. Glover
Kevin J. Mitchell
Geneva, NY



譯者簡介

方嘉德

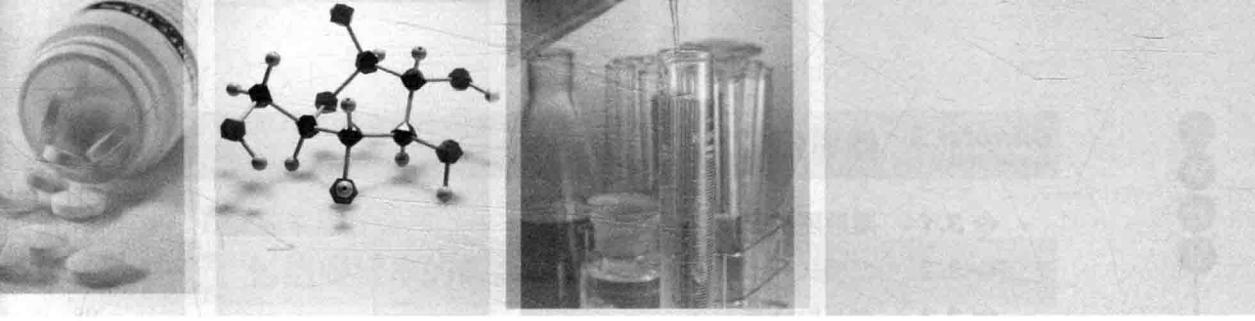
現　　職：嘉南藥理科技大學藥學系副教授

學　　歷：國立台灣大學化學系研究所博士

經　　歷：國立台灣大學化學系分析助教
海軍軍官學校理電系助教
嘉南藥專應化科副教授
嘉南藥理學院藥理系副教授
中央標準局專利外審委員
高雄市環保局專案外審委員
藥物食品分析期刊編輯委員會外審委員

研究領城：光譜分析、電分析化學、分離化學、微量分析、物種分析、數值分析。近年來從事於蛋白質體學之 LC/MS 蛋白質胺基酸定序分析與 LC/MS 微量藥物分析等研究。

譯　　作：《分析化學》(第七版)
《基礎分析化學》(第七、八、九版)
《儀器分析》(第五版)
《化學》
《分折化學》(Hage & Carr)
《生物統計》(第二版)
《生命技術概論》



目錄 CONTENTS

CHAPTER

1

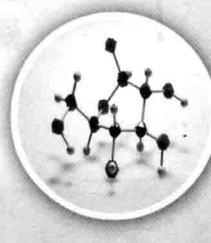
資料分析之介紹

Introduction to Data Analysis

本章大綱

- 科學方法與統計分析法
- 參數：母體之敘述性特徵值
- 統計量：樣本之敘述性特徵值
- 變數類型：連續型、離散型、序位型與類別型
- 集中趨勢之測量值：平均值、中位數與眾數
- 離中趨勢之測量值：全距、變異數與標準偏差
- 頻率數據之敘述性統計量
- 編碼對敘述統計量的影響效應
- 表格與圖形
- 四分位數和盒形圖
- 準確度、精密度和 30–300 法則



1.1 緒論

目前的生命科學研究工作包括有實驗、蒐集數據和詮釋資料等。在本教科書中將介紹進行這些基礎研究工作時所使用的一些方法。

在所有的科學領域中，都會對所謂**科學方法** (scientific method) 的實驗，進行其設計和評估工作，然而在許多研究中，這些工作時常是隱晦而不會被明確地概述之。科學方法的構成要素包括有觀測、潛在之問題或議題的系統化表達、建構假說、然後做出預測，再設計實驗以檢定所做出的預測。接著，就簡短地探討這些構成要素。

特殊事件之觀測

一般上，觀測結果可以被分類為定量的或定性的。定量的觀測結果乃是根據一些測量值而來的，譬如長度、重量、溫度和 pH 值。而定性的觀察結果則是根據能夠反映出被觀測事件的性質或特性等的一些類別為基礎，譬如男性相對於女性、罹病的相對於健康的和突變型相對於自然類型。

問題之表達

由一系列的觀測結果，通常能夠讓特定議題或未被回答的問題被系統化的表達出來。這個過程通常會採用「為什麼」問題的形式，並且暗示著其中的起因和影響關係。舉例而言，假設調查偏遠的斐濟島群眾，就能夠瞭解到大多數的成人罹患有高血壓症狀 (是指收縮壓超過 165 mmHg，而舒張壓超過 95 mmHg 的不正常血壓)。請注意！在此時，每一個單獨觀測結果是定量數據，而罹患高血壓症的百分比率則是根據樣本的定性評估為基礎的。根據初步的這些觀測結果，可能就能夠表達出一個問題：在這個母體中，為什麼有如此多的成人會罹患高血壓症？

假說之建構

假說是由觀察結果所做出的一個嘗試性解釋。良好的假說會暗示著某種起因和影響關係，並且是可以接受檢定的。

由斐濟島的群眾說明著，因為飲食、生活習慣、遺傳基因，或者這些因素的某種結合關係，可能會導致罹患高血壓症。因為我們已經注意到在他們的飲食中食用了非常大量的章魚，而且也知道章魚有很高的膽固醇含量，因

而我們就可以假設**因為飲食習慣而造成高血壓症的普及化**。

做出預測

如果已經適當地建構出假說，就應該能夠使用它來做出預測。預測是以推論的理由為基礎，並且採用「如果－則」的陳述方式。舉例而言，根據上述的假說為基礎，可以做出一個的好預測將會是：**如果高血壓症是由一種高膽固醇飲食方式所引起的，則將飲食方式改變成低膽固醇含量，就應該會降低高血壓症的罹患率**。

有效的 (被適當陳述的) 預測其判斷標準為

1. 使用「如果」子句來陳述出假說。
2. 使用「則」子句來：
   (a) 建議改變在假說中的起因 (改變飲食方式)；
   (b) 預測結果 (降低高血壓症罹患率)；
   (c) 提供進行實驗的根據。

實驗之設計

實驗的整體目的和設計都是為了完成一個目標，亦即，要檢定假說。而實驗藉由檢視因假說而產生之預測的正確或錯誤，就能夠對假說進行檢定。理論上，實驗應該只能改變或檢視由預測所建議的因素，而其他的所有因素都要維持固定不變。

在高血壓症的母體中，要如何設計出檢定飲食方式假說之實驗呢？

要檢定上述假說的最佳方法是建立一個控制組實驗。這可能包括從群眾中隨機選擇出兩組成年人，然後除了待檢驗之因素以外，對待這兩組成員的方式都應該要完全相同。呈現出所有因素的控制組代表著「正常」狀態，並且做為比較的標準或基準。而實驗組代表著「檢驗」狀態，並且除了已經被改變的變數之外，在這個例子中是飲食方式，也必須涵蓋所有的因素。如果攝食低膽固醇含量飲食的這一組呈現出**顯著**較低的高血壓症罹患率，則由數據就能支持這個假說。相反地，如果改變飲食方式對高血壓症罹患率並沒有任何影響效應，則應該重新建構新的假說，或對假說進行修正，並且也要重新設計實驗過程。最後，藉由引用數據與假說之間的關聯性，就能陳述以做為結論。

上述所概略提出的這些步驟可能顯得有些直率，然而它們通常需要被更深入地探索，並且也需要更有技巧地被適當應用之。

在上述範例中，如何選擇出各群組並不是一個微不足道的問題。它們一定要在沒有偏見的條件下被選擇出來，而且被選取的人數一定要多到足以讓研究人員對分析結果產生一個可被接受的可信水準。除此之外，變化量要達到多大才能夠支持假說呢？而所謂**統計學的顯著性**可能會不同於**生物學上的顯著性**。

統計方法的基礎是協助設計實驗，並適當地詮釋實驗結果。而統計學領域則被廣義地定義為收集、分類、總結與解析數據等各種方法和過程，以及利用數據來檢定科學性假說。**統計學** (*statistics*) 是源自於拉丁文的「國家」一詞，而且原本是指將各種普查所獲得的資料加以數值化整理，以便於描述出國家的各種形勢，舉例而言，如每年收穫的小麥浦式耳量，或達到入伍年齡的男人數量。隨著時間的經過，統計學已經變成根據自然現象，而對數值化資料所進行的科學性研究工作。當統計學被應用於生命科學時，通常就被稱為**生物統計學** (biostatistics) 或**生物計量學** (biometry)。生物統計學的基礎可以回顧到數百年前，但是生物系統的統計分析則是開始於 19 世紀初當生物學變得更為定量性和更為實驗性的年代。

1.2 母體和樣本

在今日，當面對大多數現實世界之問題中所存在的不確定性時，我們會使用統計學做為報告決策過程的一種方法。而當母體太大或難以徹底進行調查時，通常還是會希望能夠做出母體的相關歸納。在這些情況中，就會對母體進行抽樣調查，再使用樣本的一些特徵來推論出較大母體的相關特徵。這種作法可以參考圖 1.1。

真實世界的問題會牽涉到必須做出推論的那個大群體或**母體** (populations) (在相同海星物種的兩種色彩變種之間是否存在有大小尺寸的差異性？在果蠅的某種雜交子代中，正常相對於盲眼的比率是否是 3 : 1？)。有時對母體的某種特徵會感到特別有興趣 (收縮壓、重量克數、休息時的體溫)。在母體中，這些特徵的數值將會隨著不同個體而有所變化。因為這些特徵會以無法預測的方式，或者呈現為或假設為隨機改變的方式等各種方式而變化著，所以它們被稱為是**隨機變數** (random variables)。我們將會在 1.3 節中描述出各種變數類型。

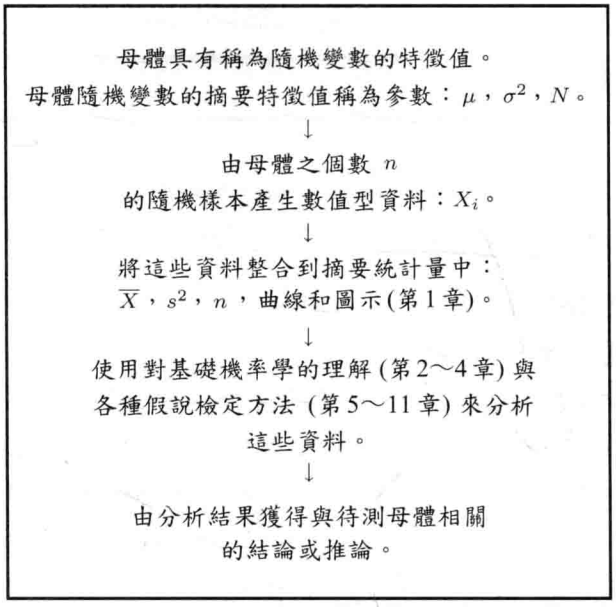


圖 1.1 統計分析的一般性作法。

當考慮到整個母體時，與隨機變數相關的**敘述性測量值**就被稱為**參數** (parameter)。其範例包括有所有綠蠵龜，*Chelonia mydas*，的平均重量，或者所有澳大利亞虎蛇，*Notechis scutatus*，的一窩蛋數目之變異數。一般上，因為母體相當大，或者要對它進行徹底調查時所費不貲，因此如果不是不可能的話，則這些參數也是難以被測量到的。所以，就被迫於對母體的子集合或**樣本** (sample) 進行檢測，再根據這個樣本對整個母體做出推論。與**樣本**之隨機變數相關的敘述性測量值被稱為**統計量** (statistic)。在蒼鷺島下蛋的 25 隻母綠蠵龜的平均重量，或在昆士蘭東南部所收集到的 50 窩澳大利亞虎蛇蛋之每一窩蛋數變化量等，都是統計量的例子。

當這些統計量不等於母體參數時，通常會希望它們和母體參數是足以接近到能夠被應用，或者可以定量出其中所潛藏的誤差值。而樣本統計量配合對機率的理解，就會構成做出與母體參數相關之推論的基礎。這種說法可以回顧參考圖 1.1。

在第 1 章中將提供整合樣本數據的一些技術。而在第 2～4 章中則將探討到一些必需的機率觀念，接著在後續的章節中將會針對假說所產生的各種預測，概述檢定這些預測所需要的各種不同技術。

1.3 變數或資料的種類

在統計學中會出現數種資料類型。每一種統計檢定方法都會要求被分析的資料必須是某種特定類型。以下將表示出最常見的變數類型。

1. 可以將**定量型變數** (quantitative variables) 區分成兩種主要分類：
   (a) 可以是實數的一些區間 (可能會是沒有邊界的) 中之任何數值的**連續變數** (continuous variables) 或**間距數據** (interval data)。常見的範例有長度、重量、溫度、體積和高度。它們是經由測量過程而獲得的。
   (b) 只能是獨立數值的**離散變數** (discrete variables)。其範例有一窩蛋的數量、每公頃中的樹木數量、每隻海星的觸足數，或者每個樣方中的品項數。它們是經由計數過程而獲得的。

2. **序位 (順序) 變數** [ranked (ordinal) variables] 是無法被測量的，然而卻有著一種自然次序。舉例而言，政治公職候選人會被個別選民予以排序。或者根據身高而由最矮到最高地安排學生，並且這種排列方式甚至是不需要進行測量的。排序數值除了所表示出的「次序」之外，並不具有其他的潛在意義。亦即，排序 2 的候選人並不會比排序 1 的候選人還要優秀兩倍 (將這個結果與量測型變數比較時，2 英尺高的樹會比 1 英尺高的樹高達兩倍之多。對於量測型變數而言，這種比率關係是有其意義的，而對於順序型變數，則不具有任何意義)。

3. **類別型數據** (categorical data) 是定性型數據。其中的一些範例有物種別、性別、遺傳基因型、表現型、健康的/罹病的以及婚姻狀態。不同於序位型數據，當它們被指定為那些分類時，並不具有「自然的」順序關係。

無論當量測型變數是從母體或樣本中被收集到的，其數值必須能夠以某些方法來摘錄或整合。元素母體的摘要敘述性特徵值被稱為**母體參數** (population parameters) 或只稱之為**參數** (parameters)。參數的計算過程則需要對母體**每一個**成員的量測型變數值有所瞭解。通常會以希臘字母來表示這些參數，而且在整個母體中，它們的數值並不會發生改變。元素樣本，亦即母體的一個子集合，之摘要敘述性特徵值被稱為**統計量** (statistics)。而隨著母體樣本的選擇結果不同，樣本統計量會呈現出不同的數值。統計量會以各種符號