

# 審定序

統計學是一門集合了資料蒐集、資料處理和資料分析的學科。通過合理的統計計算，可以從大量的資料中篩選出有效的資訊。透過統計分析軟體的開發與使用，讓統計學更廣為人所接受，特別是實證研究工作者間更是如此。統計工具的種類很多，例如社會科學領域中常見的SPSS軟體，它具有操作簡易、功能齊全、結果輸出完善等優點。此外，Excel作為資料表格軟體，也有一定的統計計算功能，包括圖示統計和函數計算功能。但這些軟體有一種共同的特徵，那就是必須付費使用。

R語言是在1980年代後期由AT&T實驗室開發，最大的特點是完全免費、資源公開的語言系統。R語言是從S語言中發展起來的一種新型統計軟體，可以在Linux、Window和Mac環境運行。與傳統的S語言相比，R語言具有如下的一些特點。首先，R語言是一種免費的統計軟體，可以從互聯網上進行免費的下載。其次，R語言具有多種統計程式，具有強大的統計功能，並具有輔助的繪圖功能，可以將相關的資料通過圖形的方式展現出來。

與其它統計軟體如SPSS、SAS等相比，R語言的特點是可編程。作為一個開放的統計程式軟體，它的語法通俗易懂，讓大部分初學者容易學會且掌握其語法。本書《R語言與資料分析實戰》以R語言的「程式設計屬性」為中心，內容涵蓋R語言基礎理論到實際資料分析，通過分析模型和演算法等更實用的範例，講解了資料視覺化、統計分析、資料採擷、機器學習等常用的方法。同時書中還收錄了作者的實戰經驗和學習體會，可以解決資料統計分析過程中出現的各種問題。如果你是一名資料統計分析的研究人員，本書將是一本不可多得的參考

書：深化理解與認識R軟體的應用，進一步提高資料統計分析水準。

吳政達

2019年2月

# 前言（作者序）

## 資料分析的起點——R程式設計！

網路Web、手機App（Mobile App）、社群網絡（Social Networking）、搜尋引擎（Search）、大資料（Big Data）是貫穿當今時代的關鍵字，將其串聯在一起的另一個關鍵字就是基於資料分析（data analysis）與資料決策（data-based decision making）。將分析與決策應用於網頁的典型案例是美國總統B. Obama募集6000萬美元選舉資金的事情，<sup>1</sup>工作人員製作了兩種設計風格的網頁，並分析（稱為A/B測試）使用哪種設計能夠吸引更多選民。這種分析（A/B 測試）不僅可以用於App行銷，也可以用於開發App。另一個資料分析的例子是分析社交網路圖的結構，或者更改社群網站的頁面組成以觀察使用者反應。在網頁搜索中也進行過大量實驗。<sup>2</sup>最近，應用大資料進行資料分析備受青睞，分析物件甚至包含大資料系統本身如何快速運行。資料分析的下一步是預測分析（predictive analytics），它是決策的根基。

隨著分析、預測、決策等話題被提起，相信R語言接下來也會受到關注。為什麼這樣說呢？首先，R語言是一種專門語言，重點在於資料分析、統計分析、機器學習、資料視覺化。使用R提供的多種包能夠輕鬆解決分析與預測問題。同時，R也是一種程式設計語言，容易擴展，

- 
- 1 此事相關報導請參見<http://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-byrunning-a-simple-experiment/>
  - 2 *Overlapping Experiment Infrastructure: More, Better, Faster Experimentation*, Proceedings 16th Conference on Knowledge Discovery and Data Mining, 2010, ACM. <http://research.google.com/pubs/pub36500.html>

適用於解決多種問題。其次，R是一種自由軟體（free software），任何個人、企業、學校、機關都可以免費使用，無需背負沉重的經濟負擔。第三，R背後有強大的社群（community），社群中開發的多種分析包都是免費提供的。最後，R有許多的相關輔助資料且圖書紛紛被出版。現在，幾乎任何一本統計分析圖書都使用R語言編寫示例程式碼。畢竟，親自動手編寫並運程式碼與只使用筆紙學習分析方法有著很大不同。

人們對R語言學習熱情的高漲促使了本書的誕生與出版。本書是筆者在多年學習筆記的基礎上編寫而成的，這些筆記是筆者為了使用R進行機器學習而整理的，在筆者的個人部落格（<http://mkseo.pe.kr/stats>）上可以看到。整理並挑選示例時，筆者參考了多種圖書與資料，每當遇到問題，筆者都會使用搜尋引擎和StackOverflow（<http://stackoverflow.com>）尋找答案。隨著資料的增加，逐漸形成了圖書的形態，最後促使本書產生。書中整理了大量R初學者經常遇到的問題及答案。透過閱讀本書，讀者可以輕鬆學習R語言並掌握應用方法，不必再經歷筆者當時學習的痛苦了。

本書韓文版的順利出版得益於Gilbut出版社한동훈課長和서형철組長的幫助，신경근先生幫我確定了全書的行文風格與方向。此外，還要感謝Gilbut出版社的相關工作人員，他們為本書的出版付出了巨大努力。

感謝我的妻子。對於每個週末都要坐在電腦前的丈夫，她心裡不免會有些怨言，但從未流露出來，也從未說出口，只是一直陪在我身邊默默等待。謝謝你的鼓勵！

最後，感謝購買本書的讀者朋友們。寫作本書時，筆者已竭盡所能，傾注大量心血，但由於自身的不足，難免會出現各類問題。如果大家有在閱讀過程中發現任何問題，請給發送郵件給筆者（[minkoo](mailto:minkoo)。

seo@gmail.com)，筆者將盡自己所能為您解答。謝謝！

徐珉久

2014年10月