はじめに

1989 年に筆者が本格的に利用したデータ解析環境は，UNIX 上で稼働していた S システムである．1990 年前後は，専用回線によってインターネットに接続されたワークステーションが大学などの教育・研究機関に導入され始めた時期であり，そのような環境は当時としては非常に高性能で魅力的なものであったが，個人で占有して利用することは価格等の面で夢であったことを記憶している．しかしながら，四半世紀あまりを経た現在，インターネット・ハードウェア・ソフトウェアを利用する上での「価格」,「利便性」,「性能」，そして「情報量」などの点で，当初の想像を遙かに超えた発展があったことには驚きを禁じ得ない．特にソフトウェアに関しては，本書で扱っているデータ解析環境 R[1)]が，これらの観点を満たすものの象徴といえよう．本書は，このような劇的な改善があった情報通信環境のもとで R を利用し，その設計のベースとなっている「探索的データ解析」における要約と可視化の重要性を再認識しながら，「統計学の基礎」を学ぶことを目的としている．

本書は，「目的別」に R に関する事項を学ぶために，R 自体を学ぶ部分 (第 I 部) と統計学の基礎を学ぶ部分 (第 II 部)，そして，実際に R を使ってデータ解析と統計的推測を行う部分 (第 III 部) に分ける構成とした．

まず，第 I 部では，初学者が最初に疑問をもつであろう「R とは何か」ということに対する解説からはじめて，R を利用する上で重要となる情報のソースや入手法などを述べた．また，利用をはじめるときの設定や利用する際に重要となる各種の用語 (オブジェクト，演算，関数) と，それらの役割や利用法について述べた．さらに，実際にデータ解析を行う際に必須となる R へのデータの読み込み方や，データを可視化するために必須となる R におけるグラフィック環境について，最近の話題と共に述べた．また，R を利用するに従って自然に必要となる，新たな関数の定義の方法についても与えた．

次に，第 II 部では，まず確率変数と確率分布を多変量の場合も含めて述べた後，母集団分布と標本分布について解説した．通常の統計学の教科書では，データを要約したり可視化したりする「データ処理」について，確率分布などの事項よりも前に述べていることが多いが，本書では，特に「分布」の概念を先に述べることによって，実際のデータの要約と可視化を行う目的が，分布の情報をデータから得ることであることを理解していただけるようにした．また，確率に関する事項は，確率変数や確率分布の基礎となっているために，通常の統計学の教科書でははじめの章で取り上げられることが多いが，本書では統計学を R で学ぶということを主目的としたため，その重要性も理解した上で，付録 A として与えた．

最後に，第 III 部では，実際のデータを要約し，可視化する方法を実践的に述べた後，統計的推測の 3 本柱である「推定」・「検定」・「回帰」について，R を利用して実際のデータを利用しながら体系的に学べるような構成とした．検定については，特に 2 標本問題についても章を別立てにしており，回帰の章では，説明変数が単一の場合の「単回帰分析」と 2 個の場合の「重

1) 日本では，1990 年代の半ばから R の存在がネットニュースなどで聞かれるようになり，1999 年に筆者が講義・演習の教育現場で実際に利用を開始した当時でも，学部レベルの統計学の講義を行うための機能 (関数) が実装されていたことを思い出す．このことは，フリーの本格的なデータ解析環境の到来を予感させるものであった．

回帰分析」を扱った．通常の統計学の教科書では，単回帰の次に一般の個数の説明変数をもつ場合の重回帰が扱われるが，本書では，ステップ・バイ・ステップで回帰分析を学べるように配慮した．また，データの要約・可視化の結果を回帰を行うときの知見を得るための情報として利用しており，このことは，「探索的データ解析」の立場そのものであることに注意しよう．

なお，付録 B では R の tips[2] として，R の実行環境を調べる方法やスクリプトの管理，R のグラフィック機能について述べた．また，コマンド・ライン・エディタの利用法，Windows 環境下での目的別ショートカットの作成法，パッケージのインストール法や R の統合開発環境である RStudio に関する話題にもふれた．そして，付録 C では，正規分布と標本分布に関する上側確率や上側パーセント点に関する値を R を用いて求める方法についてまとめた．さらに，付録 D では，本書で利用したデータの解説を，付録 E では，本書で独自に用意した関数のソースコードを与えた[3]．

なお，本書で引用したインターネット上のリンク情報などは，原稿執筆時点でのものであること，また，本書では R version 3.4.3 (2017-11-30) を利用していることに注意していただきたい．

これまでの R の発展は，多くの開発者たちによってなされたたゆまぬ努力の成果であることは疑いのないことであるが，R のオリジナル開発者である Ross Ihaka 氏と Robert Gentleman 氏に感謝の意を表したい．特に，Ross Ihaka 氏には 2003 年から 2004 年の 1 年間，彼が在籍するオークランド大学統計学科に滞在する機会を与えていただき，数多くの貴重なコメントをいただいたことに感謝したい．

また，本書は，これまで大阪大学，関西学院大学，大阪府立大学，神戸大学における講義で利用した資料に大幅に加筆・修正を加えて作成したものである．その間，多くの学生から質問・意見などをいただいたことに感謝したい．特に，神戸大学発達科学部の鷲見菜月氏と関西学院大学商学部の荒川聖悟氏，柳 麻衣氏には本稿のドラフトを読んでいただき，数多くの有益なコメントをいただいた．また，神戸大学の阪本雄二氏からは本書の構成や内容に関する貴重なご指摘をいただいた．ここに記して感謝の意を表したい．

末筆となったが，本書のドラフトに関する製本などの雑務を毎回快く引き受けていただいた関西学院大学商学部研究資料室の高瀬 忍氏と，本書を世に出すことに粘り強い対応と多大なる励ましをいただいた裳華房の小野達也氏に心から御礼を申し上げる．

2018 年　初秋

地 道 正 行

---

2)　情報環境利用上の便利な技法などのこと．

3)　本書で扱うデータと関数は，裳華房のホームページにある本書に関する Web ページからダウンロード可能である．ただし，本書で利用している「新生児の身体測定データ」と「企業の財務データ」の二次利用は許可されていないことにご注意いただきたい．なお，このサイトでは，データの「分散」を「2 乗平均」から「平均の 2 乗」を引くことによって求めることができるという，いわゆる「分散公式」の幾何学的説明と，説明変数の個数が一般の場合の重回帰分析について述べた補足資料 (補遺) も掲載しているので，併せて利用してほしい．