本シリーズの編集にあたって

社会の進化に伴い，統計科学の環境が大きく変化している．その主な変化として次のような点があげられる．1) データの収集の方法が多様化されている．2) データの平均サイズがますます大きくなっている．3) データの流通が容易になっている．4) 統計計算やシミュレーションに必要となるコンピュータがますます安価になっている．5) 統計計算やシミュレーションの専用ソフトが無料で入手可能になった．6) 統計科学の役割の重要性の認知度が向上している．

このようなさまざまな変化は，統計的データ解析の新しい手法の開発と応用を促し，データマイニング (data mining) や統計的機械学習 (statistical machine learning) のような新しい研究分野が生まれるようになり，その応用が急速に広がっている．従来の統計学，近年のデータマイニングや機械学習（マシンラーニング）に関する定義はいろいろあるが，共通点はデータを対象としていることであるので，本シリーズではこれらを包含する用語として，狭義のデータサイエンス (data science) を用いることにする．

データサイエンスは，広義ではデータの収集，加工，蓄積，管理，流通，解析，マイニングなど，データの流れの上流から下流までを貫く科学である．昨今，データサイエンスは，工学，医学，薬学，生命科学，社会科学（社会，経済，マーケティングなど），心理学，教育学はもちろんのこと，文化学のような，従来は統計学やデータ解析があまり応用されていなかった分野でも，データサイエンスの手法による斬新な研究成果が多く報告されている．データサイエンスは，あらゆる分野において必要となる万人の科学と言っても過言ではない．

データ解析の手法のほとんどは数理的理論に基づいて開発されているので，データサイエンスに関する解説書では，数式を避けると厳密な説明ができなくなる．非理工系の研究者の中には数式が苦手である方が多いため，非理工系の研究分野におけるデータサイエンスの適用が遅れている．一方，データ解析のツールを用いると，数理的な理論が分からなくても，データを入力すると何らかの結果が出力され，形式上はデータ解析が可能な時代になっている．しかし，データ解析の理論に関する理解が不十分であると，統計手法の利用を間違えたり，出力された結果の解析を誤ったりする可能性がある．

データ解析を行うには，用いる手法の数理的理論の理解だけではなく，ツールを用いてデータを解析しなければならない．そのためには，データサイエンスの基礎理論を理解した上でツールを用いてデータを操作し，データ解析やデータマイニングを行うことが望ましい．データ解析やデータマイニングの手法は，データの構造と目的に依存する．万能なデータ解析やマイニングの

手法はない．データ解析やマイニングを行う際には，データの構造や目的に合う手法を用いることが必要である．そのためには，用いる手法の理論を正しく理解することが必要である．

データ解析の手軽なソフトとしては，表計算ソフト Excel や Calc がある．前者はマイクロソフト社の有料ソフトであり，後者はサン・マイクロシステムズ社が開発したフリーソフトである．最近，個人，法人を問わず，ほとんどのパソコンには Excel がインストールされていることもあり，広く利用されている．表計算ソフトは，データの整理や簡単な計算には便利なツールであるが，高度なデータ解析を行うためには，プログラムを作成するか追加ソフトを用いることが必要である．また，これらのソフトは列の数に制限があり，大量のデータ解析には向いていない．その一方，データ解析の専用ソフトとしては SAS，SPSS，S-PLUS などがあるが，これらは高価であるため，個人のポケットマネーでは購入しがたく，恵まれている環境でなければ使用できない．

このようなことから，1990 年代にニュージーランドのオークランド大学統計学科の Ross Ihaka とアメリカのハーバード大学の Robert Gentleman により R（R 環境，R 言語とも呼ぶ）というデータ解析ツールの開発が始められ，1997 年からは多くの賛同者が加わり，オープンソース方式で開発が続けられている．R はフリーソフトであり，インターネットが接続された環境であれば，誰でもどこでも自由にダウンロードできる．R は，基本的な統計計算の環境と専用パッケージの利用環境を提供している．2011 年の現在，公開された R 専用のフリーパッケージの数は 3 千を超えている．R では，表形式のように定型化されたデータ処理やモデリング，データマイニング，機械学習などはもちろん，定型化されていない遺伝子情報のデータ，画像データ，音声・音楽データ，テキストデータなどを解析することも可能である．R は，高度なデータ解析，繊細多様なグラフィックスの作成，データマイニング，機械学習，シミュレーションなどを行うツールであり，伝統的な統計計算やデータ解析の概念を超えたツールとして発展し続けている．従来は，研究者が考案したデータ解析の方法をエンドユーザが使用するまでには長い時間を要したが，R の普及のおかげで，研究者が考案した新しい方法をエンドユーザが使用できるようになるまでのサイクルが大幅に短縮されている．

R による統計学に関する単行本がわが国で初めて刊行されたのは 2003 年である．約 5 年の間に R に関する訳書・和書の数はすでに 30 冊を超えるようになった．これが R 普及の勢いを物語っている．しかし，その中には R による初級統計学や R のマニュアル形態のものが多く，高度なデータ解析やデータマイニングに関する理論を系統的に説明し，その方法を R で実践する，いわゆる理論と実践を両立したものが少ない．

そこで，数理的な基礎が一定程度ある方は，関連手法の数理的理論を理解し，R による実践を通じてその方法の理論と応用を学び，数理的基礎が弱い方々は，R を用いて実践的に入門し，数理的理論を徐々に理解するようにと，数理に強い，弱いに関係なく幅広く使用できる本を提供することが本シリーズの主な目的である．ただし，企画した時点ですでに上記の理念と一致する本が刊行されている分野もある．それらの内容に関しては，重複を避けている．本シリーズでは，可能であれば社会的ニーズに応じて新たな内容の巻を追加していく予定である．

各巻の著者は，それぞれの分野で教育と研究にご活躍されている専門家である．ご多忙にもかかわらずご執筆をお引き受けいただいたことに感謝する．

本シリーズがデータサイエンスの発展に少しでも寄与できれば幸いである．

編者　金 明哲

まえがき

本書は，ブートストラップ (bootstrap) 法と総称される統計的推測法の基本的な考え方と使い方を系統的に解説した本邦初の入門書である．ブートストラップ法とは，観測したデータからのサンプリング（標本抽出）により擬似データセットを生成する代表的なリサンプリング（resampling, 標本再抽出）法であり，複雑な数式を知らなくても，計算機による繰り返し計算により推定量の偏りや分散をはじめとする種々の統計的誤差の推定や統計量の分布の推定などを行える計算機指向型統計手法である．ブートストラップ法の魅力の 1 つは，複雑な理論や数式に基づく解析を，計算機を用いた大量の反復計算に置き換えて実行できることにある．本書では特に，実用上有用な母集団分布が未知の場合のブートストラップ法を中心として解説し，ブートストラップ法および R をうまく使うとどのような情報抽出が可能となるのかを具体例・計算例とともに解説する．

本書の内容は次のとおりである．まず，第 1 章では R の使用法を簡単にまとめ，データ解析においてきわめて重要な役割を果たす作図や各種の統計量の計算方法などについて，本書で使用する方法を中心として概観する．次に，第 2 章から第 4 章ではブートストラップ法の基本的な考え方を説明する．ブートストラップ法が適用可能な問題として，第 2 章では統計量の偏りや分散の推定，およびパラメータに対する信頼区間の構成を例として取り上げ，ブートストラップ法の基本的な考え方と推定アルゴリズムを述べる．第 3 章と第 4 章では，それぞれの方法を詳述している．

第 5 章以降は，ブートストラップ法の適用に関する応用的な側面を概観する．第 5 章では，ブートストラップ検定，および関連するモンテカルロ検定と並べ替え検定について紹介している．第 6 章では，ブートストラップ法に基づく回帰分析法を概観する．回帰係数の推定量の分布や精度の推定，信頼区間の構成といった問題はブートストラップ法の得意分野であり，第 2 章から第 4 章で解説した方法がどのように応用されるかについて紹介する．第 7 章と第 8 章では，より発展的な話題を扱っている．第 7 章では，時系列解析にブートストラップ法を適用する方法をまとめている．第 8 章では，効率的リサンプリング法と総称される方法を紹介する．この場合の効率的とは，何も工夫しない場合（データからの無作為抽出）と比較して，リサンプリング回数を減らせる，あるいはシミュレーションの時間を短縮できるという意味である．このようなことは，大量のデータを扱う場合には重要であろう．

ブートストラップ法の発展・普及には，Efron and Tibshirani (1993) と Davison and Hinkley (1997) の 2 冊の成書が大きな役割を果たしてきた．本書はこの 2 冊の標準的なテキストの中間的なレベルを目標とし，統計学の初学者だけではなく，実務家や大学院生，また研究者にとって

も参考になるように心がけて執筆した．また本書では，現実のデータ解析などに応用できるように，推定や検定を行うための具体的な計算手順をできるだけまとめ，また，対応する R のプログラム（ソースコード）も掲載するようにした．ただし，簡潔さよりも計算手順のわかりやすさを優先して R のプログラムを記述した部分もある．本書執筆時には，ブートストラップ法関連の Contributed Packages が 200 近くあることを確認しており，これはブートストラップ法がさまざまな状況で使われるようになったことの現れであるといえよう．しかし，そのようなパッケージの使用は必要最低限にとどめた．それは，パッケージに含まれる関数を使用すれば，プログラムとしては簡単に記述できる部分もあるが，本書がブートストラップ法の入門書であることを考慮し，ブートストラップ法の基本的な考え方とそれに対応する計算手順の説明に重点を置いたからである．したがって，現実のデータ解析においては若干冗長と思われる記述もあるので，そのようなところは読者の方々の工夫に委ねたい．また，各章の最後には文献案内の節を配置し，各章をより深く理解するための文献を紹介しているので，参考にしていただきたい．

本書で使用した主要なパッケージは, Efron and Tibshirani (1993) および Davison and Hinkley (1997) で使用されているデータや関数などを集めた bootstrap パッケージと boot パッケージである．boot パッケージは R の実行ファイルとともにインストールされるが，bootstrap パッケージは各自でインストールが必要である．本書に掲載した R のプログラムは，Windows XP, Vista, 7 上の R version 2.13.2 で動作確認をした．

本書が完成するまでに多くの方々にお世話になった．まず，筆者たちが学生時代からお世話になっている中央大学理工学部の田栗正章客員教授に感謝の意を表したい．田栗客員教授には，サンプリング，リサンプリング理論をはじめとして，統計学のさまざまな研究分野に興味をもつきっかけを与えていただいた．本書の一部は，筆者たちが北海道大学工学部・大学院工学研究科・大学院情報科学研究科，北海学園大学工学部，千葉大学大学院理学研究科および国立保健医療科学院・生物統計分野の大学院などで担当した授業の内容をもとに執筆している．これらの授業に出席された学生諸君にも謝意を表したい．特に，千葉大学大学院理学研究科博士後期課程の湯毅平君には，本書の草稿に目を通していただき，数多くの有益なご指摘をいただいた．最後に，本書の執筆を勧めてくださった同志社大学文化情報学部の金明哲教授と共立出版編集部の横田穂波氏には，原稿の完成が予定よりも大幅に遅れたにもかかわらず，温かく見守っていただいた．この場をお借りして御礼を申し上げたい．

2011 年 10 月

汪 金芳，桜井 裕仁