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．2009 年の現在，公開された R 専用のフリーパッケージの数は 2 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

現代の情報化社会では，観測方法や測定技術の発達によって，さまざまなデータが大量に取得・蓄積されている．それに伴い，大量のデータからの有用な情報抽出や本質的な知識獲得の技術開発もますます重要性を増している．とりわけ，大規模データに基づく予測，シミュレーション，データマイニングなどデータサイエンスの方法が重要視されつつある．そこで，このようなデータサイエンスの遂行にあたり，ベイズ統計データ解析の手法が必要不可欠となる．

従来の標本理論に依拠した統計学では，サンプル・サイズが大きいほど統計的推定の精度も向上する．また，大規模データに付随した計算上の困難は，計算機器の高速化・大容量化と計算法の高度化によってかなり軽減される．したがって，大規模データの統計解析に際して，特段の問題は生じないように思われる．しかし，従来の統計手法では，システムの複雑な挙動を捉えることが困難である．すなわち，複雑な振舞いを示す自然現象や社会現象を適切に分析し難い．

例えば，経済は生きものであるといわれるように，経済構造は時間の経過とともに変化する．したがって，経済の動的システムを厳密に分析するには，複雑な構造を表現できるモデルが必要となる．こうした観点から，経済の構造パラメータが時間の経過とともに変化するモデル構築が考えられよう．ただし，そのようなモデルを構築するとデータの数に比較してパラメータ数が非常に多くなるため，従来の方法では安定的なパラメータ推定が困難である．こうした従来困難であった問題を解決するうえで，ベイズ的モデリングの手法は強力なツールとなる．ベイズ的な統計解析法では，観測データからの情報と観測データ以外の情報を統合することでパラメータの確率分布を生成し，その確率分布で統計的推測を行う．この観測データ以外の人間の知恵から得られる情報を事前情報という．

上述の例のようにパラメータの時間的変化の構造を表現したいとき，経験的知見からパラメータの緩やかな変化パターンを想定したとしよう．こうした想定は一見取るに足らないことのように思われるかもしれないが，この事前情報によって平滑化事前分布と呼ばれる事前分布が導入でき，パラメータの推定上きわめて有用である．平滑化事前分布の特徴は，パラメータ変化の滑らかさを規定する動学方程式と初期値を設定しておけば，パラメータの変化パターンがモデルの構造によって自動的に決まることである．つまり，このような事前分布に基づくモデルは自由度が適度に抑えられているため安定的なパラメータ推定ができ，しかも構造上高い柔軟性を有する．それゆえ，近年，平滑化事前分布に基づくベイズ的モデリングの手法は，動的構造をもつシステム分析の有力なアプローチとして広範な分野で適用されている．

ベイズ統計学の基本的な考え方は，1763 年に Bayes によって提唱された当初から約 250 年経っており，それ自体古い歴史をもっている．しかし，実践的な統計モデリングとして確固たる地位を築いたのは比較的最近のことである．ベイズ的な統計推論では，理論的には観測データの統計

モデル（尤度）とモデルに含まれるパラメータの事前分布さえあれば，事後分布が自動的に導かれる．ただし，事後分布が解析的に導かれるのはごく少数のケースであり，多くの場合は数値計算を必要とする．ベイズ統計におけるもう 1 つの難点は事前分布の構築である．通常，十分な論拠をもって事前分布を提案できる場合は少なく，便宜的に設定することが多い．そのため，ベイズ統計学の推論方式は恣意的であると批判されてきた．こうした難点がベイズ的方法の普及の足枷になっていたと思われる．

近年におけるベイズ的統計解析の普及の背景には 2 つの要因があるといえよう．第 1 は計算機器の高速化・大容量化と計算法の高度化である．そして，第 2 は統計モデリングの考え方とモデル評価方法をコアとする情報量統計学の誕生である．従来の数理統計学では，モデルは真の分布そのものであると仮定しているので，いかなる事前分布も統計解析に入る余地がない．しかし，情報量統計学における統計モデリングの考え方を援用すれば，統計モデルは真の分布の近似であって，評価方法をとおして「より良い」モデルを見つけることができる．

ベイズ統計解析では，とくにモデリングの技法とパラメータ推定に関する計算法の占める比重が高く，R との相性が非常によい．著者は，本書を通読することで，ベイズ統計学の基本に関する理解が深まるよう配慮したつもりである．また，本書は応用の側面も重視しており，分析方法の解説と併せて R によるプログラムを提示する．具体的には，まず，ベイズモデルの基本概念，ベイズ型線形モデルの手法，MCMC などモンテカルロの手法，ナイーブベイズ分類器による判別分析，状態空間モデルを説明する．そして，経済時系列の季節調整，時変係数 AR モデルおよび時変係数 VAR モデルなどの状態空間モデルによる時系列解析法を解説し，R で編成したプログラムを紹介する．さらに，応用例として，時変構造をもつ生産関数モデルの構築，ヒューマンインタフェースのパフォーマンス評価のためのベイズ型モデルなどを取りあげる．なお，紙数の制限により，統計学の基礎理論・方法と R の基本に関する解説は必要最小限に留めている．

ここで謝辞を述べておきたい．まず，大学院在学時の指導教授である田邉國士氏（統計数理研究所名誉教授，早稲田大学教授），北川源四郎氏（統計数理研究所長）には，日頃から時系列解析，ベイズモデルの構築や統計計算法などに関してご指導いただいている．この場を借りて御礼を申しあげる．金明哲氏（同志社大学教授）には，日頃から公私にわたって励ましをいただいており，とりわけ本書の執筆を薦めていただいたことに謝意を表したい．野田英雄氏（山形大学准教授）には，共同研究をとおして有意義な議論をしていただいたこと，また本書の草稿を精読し，丁寧に校正していただいたことに深く謝意を表す．なお，本書の第 12 章で取りあげた CES 生産関数のモデルに基づく技術進歩のベイズ的統計解析法も野田英雄氏との共同研究の成果として得られたものである．第 13 章で取りあげたヒューマンインタフェースの性能評価の一部分は任向實氏（高知工科大学教授）との共同研究の結果を引用している．また，任向實氏には本書の応用例として使用したポインティングタスクのデータを快く提供していただいた．中村永友氏（札幌学院大学教授）には貴重なコメントを多くいただいた．樋口聖哉氏（岩手大学大学院連合農学研究科博士課程在学）には本文中の誤りをご指摘いただいた．本書の作成に協力して下さった帯広畜産大学農業経済学分野の同僚の各氏にも感謝する．また，共立出版の横田穂波氏には，執筆が予定よりも遅れてしまったにもかかわらず，さまざまな便宜をはかっていただいた．この場を借りて感謝申しあげる．最後に，本書が家族の協力の下に完成できたことを記しておく．お世話になったすべての方々に心から感謝したい．

2010 年 5 月

姜　興起