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

我々が暮らしている人間社会において，複雑な実態を把握したり，さまざまな問題を見つけたりすること，ひいては社会のあるべき仕組みを構想することにとっては，社会調査が不可欠な情報を収集・分析するための重要な方法の 1 つである．

社会の多元化，情報化，グローバル化にともない，地方，地域，国の境を越えた，さまざまな情報が新聞，雑誌，テレビ，インターネットなどにより溢れるほど発信されている．しかし，発信された情報がどこまで正しいか，またはそれは社会の真の姿を映しているかを，実際の社会調査により確かめることが必ず必要となっている．言い換えれば，社会調査は変動の激しい情報発信環境とともに，社会の様相，そして各構成員の本当の考え方を把握するための手段として発展してきている．日本における社会調査の普及をうけ，その質的な改善と，新たな人材の育成を図るため，2003 年に社会調査士資格認定機構（2008 年に社会調査協会に改称）が発足し，「社会調査士」および「専門社会調査士」資格認定が行われるようになった．

社会調査のうち，あらかじめ立てられた仮説が正しいかどうかを検証する「仮説検証型アプローチ」もあれば，これまでわからなかった事実を発見しようとする「事実発見型アプローチ」もある．両者に共通するのは調査で得られたデータを用いて物語ることである．これは今日のほとんどの社会調査の特徴でもある．したがって，世論調査や市場調査を担う実務者のみならず，学術調査を活用する研究者も，多様な社会調査データ解析を行う機会が多くなっている．

社会調査で取り扱うデータのうち，間隔尺度や比率尺度で測った量的データより，名義尺度や順序尺度で測った質的データが圧倒的に多いので，これまで出版された多くの統計関係書に取り上げられない質的データ解析の諸方法を用いることが必要となる．本書は，社会調査データによく用いられる解析方法に焦点を絞り，諸方法の概要を説明したうえで，R による具体的な分析手順および出力結果を解説する．本書の最大の狙いは，社会調査・数理心理実験の実務者・現役大学生を対象に，「データを中心に物語る」という視点から，社会調査の基本的な考え方と実践的なデータ解析方法を示すことにある．本書の主な特徴を，以下に示す．

(1) 数学や統計学の基礎知識に自信のない文系出身の読者を対象に，社会調査の基本概念，標本抽出方法および調査実施方法を中心に，予備知識として説明する．

(2) 分析例に実際の社会調査データを最大限に用いて，各種データ解析の進め方および分析結果の読み方について詳しく提示する．

(3) 読者の異なるニーズに合わせて，量的データと質的データに分けたうえで，1 つの変数，2

つの変数および多変数の解析法をそれぞれの章として組み込む.

(4) R コンソールに慣れていない読者のために，各節に取り上げているデータ解析法を説明する後に，R コマンダーによる分析手順を簡潔に解説する.

本書の構成は，社会調査の現場で役立つという立場から，「I R の概要」「II 調査データの収集・整理」「III 単一変数の要約」「IV 2 変数の記述・推測」「V 多変量の探索的解析」の 5 部に分けて，計 15 章からなっている.

第 1 章では，フリーソフト R 言語の特徴，基本操作，データの入出力・計算，パッケージなどについて説明する.

第 2 章～第 5 章の 4 章では，社会調査の仕組み，標本抽出方法の基本的な考え方および具体的な抽出手順，社会調査データの構造およびデータの加工方法などについて述べる.

第 6 章と第 7 章の 2 章では，単一量的変数の基本統計量と可視化，単一質的変数の単純集計とグラフ作成などについて記述する.

第 8 章～第 11 章の 4 章では，2 つの量的変数および質的変数の関連分析方法，統計的仮説検定およびクロス表の独立性検定などを解説する.

第 12 章～第 15 章の 4 章では，量的多変量解析方法として重回帰分析，主成分分析，因子分析およびクラスター分析，質的多変量解析方法として数量化 I 類，ロジスティック回帰，対数線形モデル，対応分析，数量化 III 類および多次元尺度の概要に及ぶ分析手順などについて示す.

本書は，社会調査の第一線で活動している実務者や研究者の参考書と想定するのと同時に，学部や大学院科目の社会調査演習，数理心理実験，データ分析方法，R 言語演習などの教科書として利用いただける．なお，本書に収録した内容は，社会調査士認定機構が定めた社会調査士資格の標準カリキュラム【E】量的データ解析の方法に関する科目，【F】質的な分析の方法に関する科目および【G】社会調査の実習を中心とする科目にも対応している.

本書の執筆を振り返ってみると，高度な知識を簡単に説明する難しさを実感しており，本書に少なからず間違いもあることと危惧しながら，これから社会調査データや心理実験データの解析に関心をもつ方々の一助となれば幸いである.

本書の執筆にあたっては，多くの方々のご支援とご激励をうけている．特に同志社大学文化情報学部「社会調査演習」担当者浦部治一郎先生，川崎廣吉先生，西倉実季先生，村上征勝先生，宿久洋先生からは，授業内容について惜しみないご指導を賜った．また，本書に収録した社会調査事例は，統計数理研究所の吉野諒三先生，東洋英和女学院大学の林文先生，帝京大学の山岡和枝先生との共同研究による成果の一部である．ここに厚くお礼申し上げる次第である.

最後に，著者のわがままな執筆計画を許し，本書の編集作業を速やかに進めてくださった共立出版の横田穂波氏に厚くお礼を申し上げる.

2011 年 晩夏

著者一同