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

現在の私たちの生活は，数値情報と密接にかかわっている．コンピュータの利用によって，気温，湿度などの気象にかかわる情報をはじめとして，さまざまな分野で測定の自動化が進んでおり，バーコードやマークシートを利用した入力システムの開発やインターネットを利用した調査方法などによって人の行動や意見に関する情報についても短時間に大量の情報を収集することが可能となっている．また，表計算ソフトやデータベースソフト，統計解析ソフトなどを利用することによって，データの処理や統計解析が多くの人々にとって身近なものとなり，これらの技能は社会人としてのリテラシーの一つとして考えられるようになってきている．

本書で取り扱うカテゴリカルデータの解析は，質問紙調査の解析においてよく用いられる手法である．質問紙調査は，回答者の行動や感情・意見などを直接問いかける方法であり，科学的な測定では困難な内容を取り上げることも可能である．たとえば，医学においては身体に関してさまざまな物理的な測定が行われ，その結果に基づいて治療方法の開発が進められている．しかし，現在ではこのような物理的な測定の結果だけでなく，よりよく生きるという意味での生活の質 (Quality of Life) も最終目標の一つとして考えられるようになってきている．しかし，カテゴリカルデータの解析については，分割表の独立性の検定に関しては多くの本でも取り上げられてはいるが，その次の段階の解析方法に関して書かれている本は意外に少ない．また，実際の解析を行う際の計算も複雑であり，統計ソフトウエアが不可欠である．本書では，フリーのソフトウエアである R を用いて解析を行うことを前提として，実際の計算方法についてはソフトウエアを用いることにして，解析手法の意味や解釈の仕方を中心に解説をすることにしている．

カテゴリカルデータの解析は質問紙調査の結果の解析に用いられることから，読者の中には理数系の人ばかりでなく，数学を苦手としている人もいるかもしれない．そのことを意識して，数式による説明だけではなく，具体的な事例を用いた解釈についてもできるだけ詳しく解説を行っている．また，R のコマンドについては，カテゴリカルデータの解析で必要なものだけを取り上げた．基本的なコマンドの説明をあまり詳しく行うと，なかなか目的の内容まで到達できないためである．コマンドの詳しい使い方やその意味については，参考文献でも挙げているように近年良書がたくさん出版されているので，そちらを参考にしてほしい．

本書では R version 2.10.0 を基本的に用いている．バージョンの違いによって多少記述や結果が異なる可能性がある．その点は，利用しているパッケージのマニュアル等を参考にしていただきたい．なお，本書で用いた R のコードや正誤表などは共立出版のウェブページ

http://www.kyoritsu-pub.co.jp/service/service.html#019218

に掲載されているので，そちらも合せて利用していただきたい．

本書を執筆するに当たり，多くの方々のご協力をいただいた．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には，内容を構成する段階からずっとフォローしていただき，さまざまな助言をいただいた．宮崎大学医学部の竹内昌平氏と塩満智子氏には早い段階での原稿を読んでいただき，内容や表現方法についてコメントをいただいた．京都大学大学院医学研究科の大森崇准教授には粗稿の段階での原稿と R のコードのチェックをしていただき，多くの有益なコメントをいただいた．また，共立出版の横田穂波氏には，原稿がはかどらず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．この場をお借りして感謝を申し上げたい．

2010 年 1 月

藤井　良宜