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

はじめに

友人や相談相手などの人間関係，取引や提携など企業間の関係，自然界における生物間の関係，私たちの体の中の神経や遺伝子の間の関係，そしてインターネットに代表されるコンピュータ・ネットワーク．これらはすべて，構成要素が何らかの関係で結びついた網の目のような構造，すなわちネットワークとして考えることができる．ネットワーク分析とは，これらのさまざまな対象を，点と線からなるネットワークとして表現し，その構造的な特徴を探る研究方法である．

ネットワーク分析はこれまで，人間関係や集団間の関係を扱う社会学，人類学，心理学などの人文社会科学，またグラフ理論と呼ばれる数学とそれを応用した情報科学やオペレーションズ・リサーチなどの工学分野で発展してきた．近年ではそれらの領域を横断し，さらに物理学や生物学の領域をも含む「ネットワーク科学」として，学問的な関心だけでなく一般的な注目も集めるようになってきている．

しかし，ネットワーク分析を体系的に学ぶ機会はまだまだ限られている．ネットワーク分析を専門的に扱う授業を開講している大学はそう多くはないだろうし，独学しようとすれば専門書や外国語の文献を読まなければならない．そのため，ネットワーク分析に興味はあるが，数学やプログラミングに苦手意識のある文系の学生や，実務でネットワーク分析を活用したいと思っている社会人にとって，ネットワーク分析は多変量解析など他のデータ分析法に比べて手を出しにくい面があるのではないだろうか．

本書は，データ分析用のフリーソフトである R を使って，ネットワーク分析の理論と実際の分析法を学ぶことを目的としている．ネットワーク分析のためのソフトウェアとしては，UCINET や Pajek といったものが著名であるが，それらは使用できる OS が限られていたり，ソフトウェア独自のデータ形式を用いなければならないといった制約もある．それに比べ，R は Windows，Mac，Linux などの主要な OS で利用可能であり，データ形式も一般的に広く用いられているものが複数利用可能である．インターフェースもすべてではないが日本語化されており，日本語の解説書や関連ウェブサイトも比較的多い．

ただ，R はコマンドによる操作を基本としているので，マウスのクリックによる操作に慣れた読者は，最初のうちは戸惑うかもしれない．しかし，R にはネットワーク分析のための関数を数多く備えたパッケージがあるので，簡単なコマンドだけで基本的なネットワーク分析ならほとんどできてしまう．さらに，R には数学や統計学のためのさまざまな関数やグラフィックス機能が備わっているので，データを加工したり，ネットワーク分析で得られた結果を統計的に分析した

り，きれいな図として出力することが可能である．

ネットワーク分析に興味があるが，まだ実際にやってみたことはないという人でも，本書を参考にしながらRの基本的操作さえ覚えてしまえば，ネットワーク分析を容易に実践し，その面白さを実感することができるだろう．また，既に他のソフトウェアを使ってネットワーク分析をしている人にも，Rのもつ自由度の高さは魅力的だろう．例えば，ネットワーク分析用ソフトには実装されていない分析手法のプログラムを自分で書くことも，Rに備わっている関数やグラフィックス機能を利用することで，比較的容易にできるだろう．

ネットワーク分析に関心のある学生，研究者，実務家の皆さんが，ネットワーク分析を始め，研究や実務に活用していくきっかけを本書が提供できるとしたら，著者にはそれに勝る喜びはない．

2009年8月

鈴木　努

本書について

本書にはRでネットワーク分析を行うためのコマンドであるRコードを多数掲載している．Rのコマンドのほとんどは，どのOSでも共通であるが，本書の記述は利用者の多いであろうWindows版での操作を前提にしている．Windows版RにはRguiというグラフィカル・ユーザ・インターフェースが備わっているが，他のOSとの共通性を優先して，本書ではコマンドによる操作を主に用いることにする．Rのインストールをはじめ基本的な操作については，付録Aで簡単に説明しているので，Rを使うのは初めてという読者は先にそちらを一読してもらいたい．

本書では，Rでネットワーク分析を行うための追加パッケージとして，社会ネットワーク分析のためのsnaパッケージと，ネットワークおよびグラフの分析のためのigraphパッケージを利用する．これらはRの標準インストールには含まれていないので，Rをインストールした後，別途インストールする必要がある．パッケージのインストール方法は付録A.2で説明している．また，追加パッケージはRを起動した後に読み込まないと使うことができない．例えば，snaパッケージを読み込むときは次のようなコマンドを入力する（行頭の>は自動的に挿入されるので打ち込む必要はない）．

```
> library(sna)
```

本書では「snaの場合」「igraphの場合」といった説明の際に，このパッケージ読み込み操作を毎回示すことはしないので，あらかじめ読み込んでおくこと．パッケージは一度読み込めば，Rを終了するまで有効である．

ネットワーク分析のためのRの操作よりも，まずネットワーク分析について知りたいという読者は，初めはRの操作法に関する部分を読み飛ばしてもよい．しかし，データ分析を学ぶ際には，簡単な例で実際に分析をしてみるのが理解の助けになるので，実際にRにコードを入力してネットワーク分析を体験することを勧める．

本書で用いたRおよび主なパッケージのバージョンは次のとおりである．

- R version 2.9.1 (R Development Core Team, 2009, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.)
- sna version 2.0–1 (Carter T. Butts, 2009, The sna Package.)
- igraph version 0.5.2–2 (Gábor Csárdi, 2009, The igraph Package.)

今後のバージョンアップによっては，仕様変更のために本書の記述と違いが生じる可能性がある．その場合はRおよび各パッケージのマニュアルを参照されたい．

ネットワーク分析で用いる数学の多くは高校数学程度の知識で理解できるものだが，ベクトルや行列などに関する基本的な知識は必要なので，読者の必要に応じて付録Bの簡単な解説を参照されたい．