はじめに

データ分析への関心が高まっています．背景には，さまざまな分野でデータの電子化と公開が進んでいることや，高性能のパソコンや解析環境が入手しやすくなったことがあります．企業や団体，あるいは個人は，さまざまな場面で決断を求められます．後悔のない決断をするためには情報が欠かせませんが，情報はまさにデータから得られるものです．さまざまな情報を取得するためのソースとなるデータが日々蓄積されています．たとえばインターネット上のサイトのアクセス数や，ブログで特定の店が言及された回数などがわかれば，そのときの流行をうかがい知ることが可能です．

本書は，データ分析（データマイニング）のためのプログラミング技法を解説した入門書です．データ分析とは，簡単にいえば統計解析のことです．読者の中には，表計算ソフトである Excel のアドイン機能を使って *t* 検定（平均値を比較する統計的手法）を行った経験がある方もいるでしょう．最近のデータ分析では方法が多様化しています．またデータとその分析結果をわかりやすく表現するためのグラフィックス技法も多数提案されています．これによりデータ分析の応用範囲が広がり，従来は統計解析の対象とされていなかったようなデータにも，これらの手法が適用され，新規な知見が得られるようになっています．

こうした新しい手法を実行するには，Excel の機能は十分ではありません．また複雑なデータを分析する場合，データのスクリーニング，すなわちデータの整形が必要となりますが，これも Excel では非常に手間がかかるでしょう．そこで注目されているのがＲというフリーの解析ソフトウェアです．Ｒは多種多様な分析手法とグラフィックス作成手法を自由に使うことのできる解析環境です．また利用可能な手法も日々増えています．このため，世界中

の研究機関や企業で急速に導入が進んでいます．

ちなみに，R は統計解析**環境**といわれます．それは R では，既存の機能をベースに，ユーザーの側で自由に拡張機能を追加できることが意図されています．R はデータ解析の幅を広げるためのベースあるいは環境なわけです．実は，R はプログラミング言語として設計されており，この言語で命令を書くことにより，R を自由にカスタマイズできるようになるのです．

人間の言葉に日本語や英語など，さまざまな言語があるように，プログラミング言語もさまざまです．いくつか例をあげると VBA や C, C++, Java, Python, Ruby, Perl, PHP などがあります．プログラミング言語には文法と語彙があり，これは言語ごとに覚える必要があります．

ところでプログラミング言語は難しいのでしょうか？　筆者はそうは思いません．むしろプログラミング言語は，人間の言葉に比べると文法や語彙が非常に単純です．例外もありません．またプログラミング言語の構造はどれも似通っており，一つの言語に精通すれば，他の言語の習得はきわめて容易になります．

プログラミング言語を学ぶには，その環境も重要です．使いやすい環境で学んだ方が効率がよいに決まっています．そこで本書では RStudio という統合開発環境 (IDE) を利用します．RStudio は，ユーザーのパソコンにインストールされている R を拡張するソフトウェアです．RStudio を利用すると，データの読み込みやプロットの保存がマウスで直感的に操作できるようになります．

なお本書に掲載されたコードおよびデータは，http://www.kyoritsu-pub.co.jp/bookdetail/9784320110298 からダウンロードすることができます．さらに本書の付録 A.2 の手順にしたがって Gihub からレポジトリ (https://github.com/ishida-m/FirstProject01.git) を取得することも可能です．できれば後者に挑戦してみて下さい．

本書で R 言語を習得し，さらに高度なデータ解析やプログラミングへステップアップされることを筆者は期待しています．

2012 年 8 月　　石田基広