竹村彰通 他　著／編集

『データサイエンス入門』

学術図書出版社、2019年２月、212ページ、2,200円＋税

日本の社会、大学などの高等教育機関でも「データサイエンス」を巡る議論が活発化、既に定着してきている。「統計学」と云うと、大学教育の中でもあまり人気のある科目ではない気がするが、「データサイエンス」は「統計学」とどこが違うのであろうか？　こうした疑問を持つ読者がいるとしたら、まずは本書の編者の一人が書いた解説「データサイエンス入門」（岩波新書）から始めるのがよかろう。より本格的には米国のビジネス界の講義録「データサイエンス講義」（Rachel Schutt, Cathy O'Niel、翻訳、オライリージャパン）などを参照しながら、本書を読むと良いだろう。本書は分かりやすく書かれている啓蒙書である。

簡単に本書の中身を紹介しよう。

第１章は「現代社会におけるデータサイエンス」であるが、世界の中で日本が「データサイエンス分野」、その前提の「統計学分野」で大きく立ち遅れている現状を指摘、「現代のそろばん」としての「データサイエンス」の役割が強調されている。

第２章は「データ分析の基礎」ではヒストグラムや基本統計量から始まり、観察研究と実験研究の違い、クロス集計、相関関係と因果関係の違い、など記述統計の基本が説明されている。

第３章は「データサイエンスの手法」ではクロス集計から始まり、回帰分析、ベイズ推論、アソシエーション、クラスタリング、決定木、ニューラルネット、機械学習とAI、などと多くの読者にとっては聞きなれないメニューがところ狭しと並んでいる。

第４章は「コンピューターを用いた分析」ではエクセル、R、Pythonというデータサイエンスでよく使われている３つの計算ソフトがデータ分析の例を用いて説明されている。

第５章は「データサイエンスの応用事例」ではマーケティング、金融、品質管理、画像処理、音声処理、医学、など応用分野における多様な事例の簡単な紹介である。最後の短い第六章では「より進んだ学習のために」と題して幾つかの文献を紹介している。

本書の特徴としては、まず読者として「統計関係者」ではなく、高校生から大学生に対する「データサイエンス」への入門が意図されていることである。本書で取り上げている話題は既に述べたように広範に及び、本書は編者が３名、その他の著書が12名という多人数である。日本初のデータサイエンス学部の関係者であることから考えると、本書は大学生が「データサイエンス学部」に入学するとまずは学ぶオムニバス授業の概要だろう。筆者が多いのは多様な話題の取り上げ方と関連している。

そうした特徴の裏返し、不可避なのだろうが、話題は網羅的、玉石混合、したがってより詳しく知ろうとすると疑問が生じる場面が少なくな

い。例えば第３章４節のアソシエーション分析では「おむつを購入する」ことと「ビールを購入する」ことの関連が分かったことがスーパーマーケット市場に関わるマーケティングの成功例らしい。本書第２章でも簡単に言及があるが、統計学で昔から知られている「見せかけの相関」に関係しないのだろうか？　入門的な授業で現代的な話題に言及しようとすると不可避、あれもこれも小冊子に盛り込むことが教育上では有効なのだろうか？

次に「データサイエンス」が「一昔前までの伝統的な統計学の教科書」と比較して重視するのが「プログラミング技術」だろう。本書では「エクセル」、「Ｒ」、「Python」という３つの技術の初歩について説明している。エクセルはビジネス界や公的統計で普通に使われているが、最近では多くの高校生が相関係数などとともに学んでいるはず、しかし両方を組み合わせた教育はなされていないようである。エクセルに比べるとフリーソフトのＲやPythonはまだまだ一般にはなじみが薄そうである。Ｒについては統計学の専門家だけではなく、例えば日本統計学会の「統計検定」でも必須の統計計算ソフトである。Pythonは機械学習（Machine Learning）の研究やビジネス実務、特に深層学習（Deep Learning）関係などでよく使われている。総じて現代の「データサイエンス」が伝統計的な統計学と異なることを理解する上ではコンピューターを使う技術の重要性が認識できるだろう。ＲやPythonを利用できれば、かなり込み入った行列計算や数値計算は線形代数や微積分を十分に理解していなくとも、とりあえず実行することは可能である。弊害もあろうが「Big-Dataの時代」においては計算ソフトに習熟することが不可欠な教養になりつつあり、本書の説明は有用と思われる。

なお残念ながら本書の分量の制約からか、観察と実験に関する因果問題、サンプリングとランダムネスなどについて十分な説明がなく、評者としては期待外れで残念であった。しかし本書はデータサイエンス体系の第一巻である。かなり遅ればせではあるが、日本で初めて「データサイエンス学部」、「データサイエンス大学院」を設置した滋賀大学が企画している教育の具体的内容が今後、充実していくことを期待したい。多くのデータサイエンティスト、統計家を育成していく上で模範的内容となる可能性があり、「データサイエンス」の適用範囲の広さを強調したいのだと推察する。なお本書から学び始める学生諸氏にとりより適切な勉学の道筋についての助言があるのが望まかった。

最後に蛇足であるが、「日本におけるデータサイエンスの現状」を見ると、数年前から「証拠にもとづく政策」（EBPM）が叫ばれている中、本年はコロナ騒ぎ、「検査数が増えたから感染者数が増加、感染率が上昇している」という話題がメディアでは毎日のように報道されている。そういえば、情報の発信元の１つである役所だが、昨年の冬には「統計不正」で騒がれた役所ではなかろうか？　「何かが変だ？」と感じるのは本評者だけであろうか？　多くの読者が新しい時代を考察する１つのきっかけとなればと考え、本書の一読を推奨する。

評者：国友　直人・くにとも　なおと
（明治大学政治経済学部特任教授、東京大学名誉教授）