

菅原琢著『データ分析読解の技術』中公新書ラクレ（2022年）

“ビッグデータ”の活用、データサイエンティストの需要の高まりなど、統計学が脚光を浴びて久しい。そのせいなのか、テレビ、ネットニュースには「〇〇が80%！」などの調査結果をもとにした記事も多くみられるようになってきた気がする。しかし、そこで使われている数値の信憑性はどうか。まして、「80%の原因は、〇〇だからである。」という分析がされているケースでは、何をどうしたらそのような結論になってしまうのか、その分析に疑問を感じるケースも少なくない。

本書はこのような「データ分析ブーム」がもたらした怪しい“分析らしきもの”とそれに基づいた誤解や偏見などについて、分析の失敗例をもとにしながらデータを正しく読み解くための実践的な視点や方法、基礎的な知識やコツを紹介してくれるものとなっている。

第1章ではデータ分析を行う際、またはそれを読む際の基本となる「因果関係」について、「テレビで交通事故のニュースが増えたのはなぜか？」という問題を解いていく過程で実践的な技術を教えてくれる。以降7つの章では、①「分析結果」を批判的に読み、②「データ」自体に潜む問題を感知し、③現象の背後にある「因果構造」を考察するという3つの側面に分け、よく目にするデータ分析の失敗のパターンを解説しながら、データ分析の技術が様々な角度から示される。取り上げられるいくつかの例は、「美容院が多い県の住民は見栄っ張り？」、「病院での血液検査から一般市民の新型コロナウイルスの感染率がわかる？」、「次点で落選した候補者は次の選挙で当選しやすい？」、「出所者が就職すれば再犯を抑制できる？」、「新聞を読むと学力がつく？」といったものである。明らかに因果関係が怪しいものから、単純に因果関係がないといいきれないようなものまであるが、本書のユニークなところは、これらの例に対する解答をわれわれ読者に求めてくることである。もちろん試験されるわけではないので、データ分析に関わっていない人から関わっている・関わりたい人、さらにはデータ分析に精通している人まで、それぞれの立ち位置で問題と向き合いながら読んでいける仕組みとなっている。また、各章の終わりには「相関係数」や「無作為抽出の方法と実際」、「生存バイアス」、「疑似相関」などについての解説があり、データ分析に必要な統計の基礎知識を得ることもできる。

そして、本書の中で最も重要な指摘は「データ分析の最高の味方は現場の知識」という点である。因果構造を探る上で、結果を左右する要因に最も詳しいのは、結果のそばにいる人、問題発生地点にいる人、つまり「現場」にいる人であり、この現場の知識をデータ分析する時、あるいはそれを読解するとき利用することの重要性が示唆されている。

この間、新型コロナウイルス感染症の影響もあり、対面や郵送式で実施してきた多くの調査はインターネットを利用したWebアンケートへと変わった。実際、Webアンケートのツールも多くなり、かなり気軽にアンケート調査を実施することが可能となっている。ただ、事前にしっかり設計しないで作成した調査をあとから活用することは難しい。本書でも再三指摘されていたことであるが、あたかも因果関係が成立しているかのように自らに都合よいデータを使ってしまうたり、数値の相関を誤解して解釈してしまわないように、データ分析の技術を向上させるとともに現場がどうなっているのかをしっかりと認識しながら調査を設計し、数値を読み解いていきたい。

（加藤 健志）