日本計量生物学会 会報第 119 号（2015 年 11 月）
シリーズ: 計量生物学の未来に向けて

データとの真摯な対話

小森 理（福井大学）

近年の情報化社会の急激な発展により，一昔前と比べデータを取り巻く環境が一変した．ビックデータに象徴されるような質的量的に複雑で規模の大きいデータを扱うことが多くなり，データ解析前に必須の作業である「データの俯瞰」も一筋縄ではいかない場合が増えてきた．また遺伝子発現量，一塩基多型，次世代シーケンサーに象徴されるようなデータでは一般に変量数が標本数を大きく上回り，漸近論を主軸とした従来型の統計手法の適用が難しい状況も日常的となってきた．そう言った意味で計量生物学を含めた統計科学全体は新たな局面を迎えつつあると言えよう．

近年我々はデータサイエンスという言葉を頻繁に聞くようになった．1960 年代ごろから使われていたようであるが，実際に注目をされるようになったのは 1997 年の C.F. Jeff Wu 教授のミシガン大学での就任演説（Statistics=Data Science?）の頃からである．統計科学の他に情報科学，人工知能，コンピュータサイエンス，機械学習といった幅広い学問を包括する分野であり，次世代の統計科学とも解釈できる．データの収集・整理のステップ，データの解析・モデル構築のステップ，そして得られた知見に基づく意思決定のステップから基本的には構成される．ここで一番重要なステップはデータの解析やその後の意思決定ではなく，手始めに行うデータの収集・整理のステップだと私は常日頃思っている．

重要という理由は主に 2 つある．データ解析の労力の 8 割以上がデータが持つ特徴の理解に費やされるべきであること，またこのデータとの対話（interactive な知的なやり取り）には忍耐力と共にデータに向き合う真摯な姿勢が必要であることの 2 点である．データ解析がうまく行く場合というのはデータの特徴を捉え，それに基づいてモデルを構築して本質をえぐり出せた時である．データの特徴を捉えるためには，そのデータがどのように収集されたか，どのような背景を持つかの十分な理解と共に，データを見る眼を具えることが必須となる．それには R, S-plus, SAS といったデータ処理ソフトの基礎的な使い方の習熟が欠かせない．これを仲介としてデータとの知的なやり取りを何度も繰り返すうちに，データの全貌が明らかとなりその後の解析方針が自然と見えてくるのである．逆に一番避けるべきデータ解析というのは，このデータ理解のための地道な作業を疎かにし，様々な統計解析手法を試すだけで終わるその場しのぎの解析である．

2 つ目の重要な要素としてデータに向き合う姿勢を挙げた．しばしば統計に関する非難中傷を聞くことがある．一例として統計はウソをつく可能性があるという非難である．データの改竄はもっての外だが，都合の良い解析結果が出るようにデータ解析を不正に歪めてしまうことがしばしば問題視されているのである．目先の利益に惑わされることなく，上記の地道な作業を厭わず，素心深考の精神で取り組んでもらいたい．データ解析のスキルはもちろんのこと，データに向き合う解析者の心構えもこれからの計量生物学も含めたデータサイエンスの発展の上では重要な要素となるであろう．

2015 年 10 月に筆者は研究所から大学へ異動となった．計量生物学を主軸にこれからも専門性を深めるとともに，専門分野にとらわれず幅広く研究に従事していきたいと思っている．また理論と応用の両方の側面を重視し，実際に役立つ研究を目指すつもりである．そしてこれからは研究だけではなく教育にも力を注ぎ，統計科学を含めたデータサイエンスの面白さを学生に伝えるとともに，データ解析に携わる者としての心構えも学生にはしっか

りと身につけてほしいと思っている．そのことが計量生物学の今後の発展にも繋がると確信している．