データサイエンスとその普及

～ビッグデータ時代における展開～

○ 林 篤裕
(九州大学 基幹教育院)

1　はじめに

「ビッグデータ」というキーワードが日常的に聞かれるようになって久しい。以前には「データマイニング/テキストマイニング」と言う用語も盛んに聞かれた。技術の進歩に伴って GPS やスマートフォン・携帯電話に内蔵された各種センサーから逐次的に種々のデータが大量に生成されることや、それらを保存するデータストレージが安価でかつ膨大になったこと等から変量・サンプルサイズ共に大きいデータの収蔵が容易になった。これらのデータをどのように活用するかという点で統計技術が注目されその一つがビッグデータとして注目されているように感じている。

一方、従来からのデータサイエンスの考え方ではビッグデータへのアプローチとは異なっている側面も持っているように思われる。本発表では両者の相違点についてその一端を議論し、今後のデータサイエンスの普及について言及する。

2　大量データの生成時代

ビッグデータの定義として明確なものはないようであるが、特徴としては「3V」と呼ばれており、つまり(1) Volume(容量)、(2) Velocity(更新頻度)、(3) Variety(多様性)の性質を持っていると言われている。例えば携帯電話や車両が移動と共に発する位置情報を利用して特定の日時における人や車の動きを把握し、混雑や渋滞への対処方策を取り、場合に依っては異変の察知や予測に利用することが提案されている。

これらビッグデータの議論に接して感じることは、滝口における水の落下のように大量に押し寄せてくるデータに対してその活用手法を展開しているように思われる。そこにはデータ提供者・収集者の思惑や分析者の意図とは無関係に多種多様な大量のデータが時々刻々と生成・到着し、それへの対応を迫られているようにも感じる。

一方、従来の統計技術においては、データの採取には事前の緻密な設計が求められ最終の分析過程までを吟味した上でデータ収集が開始されていた。それはこれまでデータの収集費用が高価であったり、測定器が未発達であったという技術的な側面もあったであろうが、データが採取される状況を精査した上でデータを能動的に採取していた。そのためにはデータの採取場所にどのような背景や生成過程が内在しているか等も吟味・調査した後の採取であった。この点は近年のビッグデータに対する取り扱いとは異なっているように感じられる。

3　初等中等教育への期待

ビッグデータの注目と呼応したかのように、平成 20 年(小学校・中学校)と平成 21 年(高等学校)

に告示された新学習指導要領では、算数・数学領域に統計的な見方や考え方に基づいた単元が取り入れられた。これらは単に公式を覚えて数値を算出するという知識暗記型の教育ではなく、データに基づいた問題解決力の育成が期待されている。言うまでもなく統計的なものの考え方は、社会を生き抜いていくために有用なスキルであり、修得することによってより豊かな生活が展開できる。

この目的を達成するために初等中等教育を担っている教員がいろいろと工夫を凝らした授業を年次進行で展開しはじめている。高校等への進学率が 98%(H25 学校基本調査)となっている現在、教員にはデータの分析に興味を持ってもらい、データに内在する構造を明らかにできる技術としての統計の面白さを知って児童・生徒に接してもらえば、統計的なものの見方や考え方が次第に定着していくことが期待できる。

4　ビッグデータ時代のデータサイエンス

このように統計の分野には現在追い風が吹いていると捉えて良いであろう。このような中で従来からのデータサイエンスはどのように展開していけば良いのであろうか。

一つには調査対象となるデータを丁寧に扱うことの再認識ではないかと考える。データの採取が容易に行える時代になったとは言え、調査の根幹はデータの質に依っていることに変わりはない。調査目的に即したデータが採取できているかを注意深く精査することはデータ分析の基本であることを知ってもらうことであろう。また、採取されたデータの背景情報を熟知することも、測定時に混入するノイズの理解だけでなく分析結果の解釈の際にも大いに理解を助けるものになると考えられる。加えて大量に生成されるデータを考慮した解析技術の開発も必要であろう。

このようにデータの取り扱いを中心に据えた分析姿勢がより質の高い調査結果を得る鍵となり、今後のデータサイエンスの浸透に寄与すると想像できる。

5　まとめに代えて

ICT 技術の発達によりデータの採取・収蔵が容易になリ、加えて可視化技術も充実期を迎えた今、統計技術は再評価されている。この期を逃すことなく広く統計技術を知ってもらい多くの人に利活用いただく環境整備が大切と考える。その一環として統計的なものの考え方を広く知ってもらうための教育も重要であり、新しい時代に即応したデータサイエンスの普及が必要であろう。また、統計的な知識の理解度を試すには、統計質保証推進協会が実施している統計検定を利用するのも一つである。

とは言え、データサイエンスの考え方は一足飛びに普及できるものではないので、今後もいろいろな場面を通じて不断に支援して行きたいと考えている。

参考文献

1) 西内 啓(2013), 統計学が最強の学問である, ダイヤモンド社.
2) 水田 正弘(2014), ビッグデータに対する計算機統計学の役割, 日本計算機統計学会第 28 回大会論文集, PP55-56.
3) 南 弘征(2013), ビッグデータ解析環境に関する考察, 日本計算機統計学会第 27 回シンポジウム論文集, PP249-252.