一統計家からみたデータサイエンス

随　筆

足立浩平*

Data Science from the Perspective of a Statistician

Key Words：Data Science, Faculty of Data Science, Statistics

データサイエンス学部の出現と実業界からの期待

データサイエンスが世界的に流行している．その流行を示すことが，日本の大学におけるデータサイエンス学部の新設ラッシュである．

本邦での最初のデータサイエンス学部は，2017年に滋賀大学に新設され，その翌年の2018年には横浜市立大学に同学部が新設された．滋賀大学のデータサイエンス学部は同大学の彦根キャンパスにあるが，そこで開催された学部新設記念のパーティにおいて，横浜市立大学の先生が披露された祝辞が今でも記憶に残る．その祝辞は「彦根と横浜に共通する重要人物は，徳川幕府大老の井伊直弼（1815～1860）である」とのトークから始まり，「幕末・維新のような歴史的変革をデータサイエンスが引き起こすのか」と想わせられた．

その後，諸大学でデータサイエンス学部の新設が続き，2023年には，女子大学としては初めてのデータサイエンス学部が，京都女子大学に新設された．その学部に，著者は，大阪大学を定年退職した翌日の2024年4月1日から勤務し，統計学に関する授業などを行っている．著者の知る限り，以上に掲げた3つの大学を含め，2024年の9月までに，日本の計8つの大学にデータサイエンス学部が新設された．そして，学部ではなく，その下位組織として，データサイエンスやそれに関連する分野の学科・コース等が新設された大学は枚挙に暇がない．

データサイエンスの特徴は，実業界からの期待であり，そのことは，諸大学のデータサイエンス学部が，多くの企業・行政と連携関係を結んでいることに見られる．例えば，著者が勤務する京都女子大学・データサイエンス学部では，2024年9月26日現在で，JR西日本，NTT西日本，オムロン，京セラを含む14の企業，および，京都府，京都市，京都市産業技術研究所，西宮市と連携している．こうした連携先の方から聞かれた注目すべきコメントの1つは，「女性だからこそ気づける目的に向けてデータ解析を活用することを，女性データサイエンティストに期待している」との見解であり，新たな勤務先での教育指針の1つを与えられたと感銘した．京都女子大学に続いて，2025年4月には，東京の大妻女子大学にデータサイエンス学部が新設される予定である．

なお，以上のトレンドに先立つ2015年に，大阪大学には数理・データ科学教育研究センター（MMDS）が創設され，著者も，大阪大学人間科学研究科に在職していた時は，同センターを兼任していた．

データサイエンスと統計学

理学部の理学や文学部の文学と同様に，データサイエンスを簡潔に定義することは容易ではない．その代わりに，データの科学の出発点と思える簡単な式を，次に掲げよう．

$$\text{データ} = \text{モデル} + \text{誤差} = f(\theta) + \text{誤差} \qquad (1)$$

ここで，モデル$=f(\theta)$は，データの発生メカニズムを表す関数であり，θは関数に含まれるパラメータ（=未知の数）の集まりを表す．私は，すべての統

* Kohei ADACHI
1958年11月生まれ
京都大学　文学部　哲学科（心理学専攻）卒業（1982年）
京都大学　博士（文学）（1998年）
現在、京都女子大学　データサイエンス学部　データサイエンス学科　教授
大阪大学　名誉教授　博士（文学）
専門／多変量解析法　行列分解　心理統計学
TEL：075-531-8011
FAX：075-531-8021
E-mail：adachik@kyoto-wu.ac.jp

計解析の基礎に (1) のモデル式があると信じている．(1) 式を書き換えた「誤差 = データー $f(\theta)$」の関数

$$\text{誤差量} = \phi\{\text{データー} f(\theta)\} \tag{2}$$

を何らかの方法で定義して，これを最小化する θ を求めることと統計解析を定式化できる．

さらに，(1) 式の左辺と同じデータに対して，

$$\text{データ} = \text{モデル}^* + \text{誤差} = f^*(\theta^*) + \text{誤差} \tag{3}$$

のように，(1) 式のモデルとは異なるモデル $f^*(\theta^*)$ があれば，(2) 式の $f(\theta)$ を $f^*(\theta^*)$ に代えた誤差量を最小化する θ^* を求める統計解析が考えられる．そうすると，「モデル $f(\theta)$ と $f^*(\theta^*)$ のいずれがより良いか」という問いが生じる．現在の統計学では，

最小化された誤差量が小さく，
　パラメータ数が少ないモデルが，より良い　(4)

ことが数学的に導出され，(4) を満たすモデルの良否の諸基準が提案されている．なお，そうした諸基準の中でも，著名なものが，日本を代表する統計学者の赤池弘次（1927−2009）が提案した Akaike's Information Criterion（AIC）である（Akaike, 1974）．

(1) は単純な式ながら，(1) 式のデータを現象，モデルを理論に置き換えれば，

$$\text{現象} = \text{理論} + \text{誤差} \tag{5}$$

となり，観測された現象を理論化しようとする実証科学を物語っている．哲学では，古くから「現象を同程度に説明できる（つまり誤差が同程度の）複数の理論があれば，理論中に用いられる概念が少ない点でより単純 (parsimonious) な理論がより良い」と論じられてきた（例えば，Hempel, 1966）．この議論の中の「理論」を「モデル」，「概念」を「パラメータ」に置き換えると，(4) 式に合致し，(4) 式を満たす AIC などの諸基準は，上記の哲学的議論の数学的実現といえる．

ここまでの段落で，データの科学の出発点と思える (1) のモデル式がすべての統計解析の基礎にあることを主張した後，複数のモデル間の良否を比較するために統計学で導出された諸基準が，実証科学の理論の良否を比較する哲学的議論の数学的実現になっていることを論じた．以上の議論を通して，統計学がデータサイエンスの支柱であることを主張したつもりでいる．

著者の来歴と専門領域

遅くなったが，データサイエンス学部に勤めて，統計学に関する授業を受け持つ著者の来歴と専門を，本節に記す．

著者は，データサイエンスからは縁遠い感のある文学部を卒業している．ただし，著者が所属したのは，文学部の中でも当時は哲学科の中の一専攻であった心理学教室であった．いわゆる文科系の中でも，経済学と並んで，心理学は統計解析を多用する分野である．卒業後すぐに心理学に関わる職に就いたが，心理学そのものよりも，むしろ，心理学のデータを分析するための統計解析法を研究開発する心理統計学（psychometrics）なる分野に興味を移し，勤めながら，心理統計学とその基礎となる数学を独学した．そのうち，学術誌に著者の論文が採択されるようになり，いつの間にか，統計学の前につく心理という形容語もとれた統計学の世界に浸って，日本計算機統計学会の会長も務めた．いわば，文科系から理科系に転身したわけである．

さて，私の専門領域は，統計学の中でも多変量解析である．多変量解析とは，複数の変数すなわち多変量のデータを変数間の関係を考慮しながら分析する統計解析法の総称であり，多変量解析の個別手法には，回帰分析・判別分析・主成分分析・クラスター分析・因子分析などが含まれる．以上の専門領域で著者が行った仕事を，随筆としてはテクニカルになりすぎる学術誌論文ではなく，主要な 3 つの拙著を通して，紹介したい．

最初の著書は，足立（2006）の「多変量データ解析法 ―心理・社会・教育系のための入門―」である．これは，副題に記した分野をはじめとした文科系の学部生を読者に想定したテキストであり，完全な数式ではなく，日常言語が混ざった式を用いて，わかりやすく多変量解析の諸方法の原理を解説したつもりである．例えば，回帰分析のモデルを，

$$\text{売上} = c + b_1 \times \text{素材} + b_2 \times \text{値段} + b_3 \times \text{デザイン} + \text{誤差} \tag{6}$$

のように，足立（2006）では例示している．(6) 式は，(1) 式の左辺「データ」をある商品の「売上」として，(1) 式の右辺にある $f(\theta)$ を「$f(c, b_1, b_2, b_3) = c + b_1 \times$ 素材 $+ b_2 \times$ 値段 $+ b_3 \times$ デザイン」と限定したもので

ある．(6) 式のように，分析後に解が与えられるパラメータ（c, b_1, b_2, b_3）以外は日常言語を使うと，式に現れる複数項の間の役割の相違が明瞭になり，読解を促す．こうした記述スタイルによって，足立（2006）は好評を得たつもりでいる．

次の拙著は，Adachi（2016）の増補改訂版であるAdachi（2020）の英文テキストである．この著書は前段の足立（2006）とは対照的に数式ベースである．著者は数学の中でも特に行列（matrix）やベクトルの数理を扱う行列代数に基づいて，研究を行ってきたので，Adachi（2016, 2020）は，行列代数を学習しながら，多変量解析の諸方法も学習できるテキストとなっている．Adachi（2016）を記した後，幸いにも研究も進展して，それらの新知見も加えて，Adachi（2020）には，Adachi（2016）に6つの章が追加されている．追加された6つの章には，行列分解型の因子分析の章や，機械学習という名のもとに統計学で進展した領域の1つであるスパース推定を使ったスパース回帰分析の章やスパース因子分析の章が含まれる．

最後の拙著は，足立・山本（2024）であり，この著書には，著者（足立）が，2010年頃から考えてきた主成分分析と因子分析の類似と相違を総括し，両分析に伴う回転法を共著者（山本）が論じている．行列代数の中でも，著者（足立）が最も大切と信じる特異値分解から始まるのが足立・山本（2024）の特徴である．

以上のような専門から，著者が京都女子大学で受け持つ授業には，多変量解析の講義・回帰分析の講義が含まれる．さらに，心理分析という名称の講義も受け持つが，これは心理学出身という著者の来歴による．ただし，講義名の心理分析とは，出会った人の心理を分析することではなく，統計解析による心理データの分析を指す．こうした授業があるのは，統計学の実証科学への応用もデータサイエンス学部では重視されることを表している．

データサイエンス流行の前に統計学への脚光あり

2番目の節で意図した「統計学がデータサイエンスの支柱である」という主張を補強する事実を，本節に記そう．それは，冒頭に記した日本の大学でのデータサイエンス学部新設ラッシュの前に，世間からの統計学への注目が始まった前史があり，それに著者も関わったことである．

著者の認識によれば，21世紀に入る前の統計学の状況は，広く理科系から文科系に渡る実証科学の諸分野でデータの統計解析は必須であるため，統計学は有用な方法論であった．ただし，主にアカデミアの研究者や一部の専門家が使う方法論という感があった．しかし，21世紀に入って，アカデミアや専門家の狭い世界を超えて，統計学の有用性が広く認識されるようになってきた．こうした動向を促進した著名な書籍に，2013年1月に発行された西内（2013）の「統計学が最強の学問である」がある．その出版の直前に，私も一員となった次の大きな事業が始まった．

統計学の教育の充実化を目指す「データに基づく課題解決型人材育成に資する統計教育質保証」という5年計画の課題が，文部科学省の大学間連携共同教育推進事業として採択された．この採択に伴い，2012年9月に，統計教育大学間連携ネットワーク（JINSE, Japanese Inter-university Network for Statistical Education）という組織が発足した（美添，2018）．発足時の連携大学は，大阪大学・東京大学・総合研究大学院大学・青山学院大学・多摩大学・立教大学・早稲田大学・同志社大学の8校であった．JINSEの事業には，上記の連携大学に加えて，日本統計学会などの統計学に関する6つの学会，および，日本経済団体連合会などの8つの団体がステークホルダーとして参画し，JINSEは大所帯であった．

連携大学・学会から選出されたJINSEの委員は，質保証委員会やカリキュラム策定委員会などの委員会に振り分けられ，著者はカリキュラム策定委員会に組み込まれた．この委員会の最初の課題は，学部を超えた統計学入門の標準的なカリキュラムの構成であり，次の課題は，個別分野のカリキュラムの構成であった．この委員会のため，著者は，2012年の11月ころから2年間ほど，2, 3カ月おきに代表校の青山学院大学の青山キャンパスに出張した．そして，特に最初の課題のために，理学部・工学部・医学部・経済学部・経営学部・社会学部・人間科学部など広い所属学部に渡る諸大学の統計学者が1つの教室に集結して，カリキュラムについて論じあったことは，大変ながらも，良い経験・思い出となっている．

カリキュラム策定委員会の仕事には，全国で統計関連科目を担当している教員等を対象とするアンケート調査も含まれ，調査の結果から，統計学を教える人が不足している状況が推察された．例えば，「ある大学の心理学科では統計法の授業を要するが，それを担当する統計学の専門家がいず，統計解析を得意とする心理学の先生が統計法の授業を担当させられる」といったケースの存在が察せられた．

5年間のJINSEの事業成果の波及効果が，美添（2018, p. 181）に列挙される．その1つが，冒頭に記した滋賀大学に始まるデータサイエンス学部の新設である．つまり，JINSEに関わって大変な思いもしたが，その波及効果として新設された学部に著者は勤務していることになる．なお，上記の波及効果には，参考文献の統計教育大学間連携ネットワーク監修（2017）の書籍「現代統計学」も含まれ，その11章のうち，多変量解析に関わる第3〜6章は著者が執筆した．

おわりに

以上に記した事を，著者の時間軸でまとめよう．著者は，心理学から統計学に転身した統計家である．その転身途上で，第2節の(1)式に，統計学に留まらず実証科学も表現できる普遍性を感じた．その後，前節に記すデータサイエンス流行の前史に立ち会った．そして，冒頭に記したデータサイエンス学部の新設ラッシュの中，今は，著者もデータサイエンス学部に勤める．その学部は女子大学にあり，女性データサイエンティストの育成が著者の主要な課題となっている．

参考文献

足立浩平（2006）多変量データ解析法 ―心理・社会・教育系のための入門―．ナカニシヤ出版．

Adachi (2016). *Matrix-based introduction to multivariate data analysis*. Springer.

Adachi (2020) *Matrix-based introduction to multivariate data analysis, Second Edition*. Springer.

足立浩平・山本倫生（2024）主成分分析と因子分析 ―特異値分解を出発点として―．共立出版．

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Hempel, C. (1966). *Philosophy in natural science*. Prentice Hall.

西内　啓（2013）．統計学が最強の学問である ―データ社会を生き抜くための武器と教養―．ダイヤモンド社

統計教育大学間連携ネットワーク（2017）．現代統計学．日本評論社

美添泰人（2018）統計教育連携ネットワーク（JINSE）の展開．統計数理，**66**, 177-186.

前はホオジロガモ後ろはオナガガモ