統計学の不易流行

仁科　健

1　はじめに

まず、本稿執筆のスタンスを述べておきたい。二つある。

特集テーマである「変わること／変わらないこと」を、統計学を対象にという原稿依頼を頂戴した。変わったものに気づくのは結構容易い。本稿のテーマである統計学であれば、「ビッグデータ」とか「データサイエンス」など、ここ数年で新聞紙上を賑わし、専門家でなくとも耳にすることばである。ＩＴ技術と計測技術の発展によって、データを獲得する環境がドラスティックに変化し、データの質／量ともに変わった。統計学にはその変化に対応した発展がみられることが容易に想像できる。したがって、この二つをキーワードとして検索

していけば、統計学が重要な世の中になってきたという認識に立ちどころにたどり着く。変わったことへの気づきは比較的容易である。

一方、変わらないものに気づくのは結構やっかいだ。仮説検定がそうである。リスクを定量的に（有意水準という）明示した上で「違いがある」ことは検証できても、「違いがない」ことを統計的に検証するのは少々やっかいである。例えば、薬効分野で副作用がないことを検証するためには、まさに「違いがない」ことを検証しなければいけない。「同等性（非劣性）の検定」という方法がコンセンサスを得たのは比較的最近である。同等性の検定は、どこまでの違いを「違わない」とするかを決めなければならない。

そこで、本稿執筆の一つ目のスタンスである。まず、変わったものを挙げておいて、その中から変わらない要素を探していこうと思う。変化を整理すると、その過程で不変な点にも気づくはずである。言わば、「統計学における不易流行」を探るというスタンスで執筆に臨みたい。

統計学の動向を論ずるとき、コンピュータの発展と切り離すことはできない。筆者もそれを実感した世代である。学生時代（一九七一年名工大入学）、コンピュータの入力媒体は紙テープであった。何しろ、学部の講義では計算尺を使っていた時代である。修士の時には、名

古屋大学の大型計算機のお世話になった。入力媒体はパンチカードである。カード読み取り機を使って入力するわけだが、しばしば、「ジャムって」しまう。演算処理はバッチ処理だった。その後 Time Sharing System（ＴＳＳ）となったが、しばらくは、端末を得るために名大通いが続いた。名工大の研究室に居ながらにして名大の大型計算機センターを利用できるようになったのは、教員になって（一九七七年）からであったと記憶している。今は、ノートパソコンで、しかもユーザーフレンドリーなデータ解析ソフトを使い、クリック一つで解析結果が得られる。R や Python などのオープンソースのフリー解析ソフトの普及も見逃せない。

そこで、二つ目のスタンスである。完全に言い訳である。統計学の不易流行を語るときに、コンピュータのハード／ソフトの動向は欠かせない。しかし、コンピュータ技術に不勉強な著者にはこれは難儀である。コンピュータ技術に言及しないことをご容赦いただきたい。

2　統計学への追い風

二〇一〇年頃から、「ビッグデータ」や「データサイエンス」ということばがメディアに登場するようになった。これまでも、データマイニングやテキストマイニングが話題になっ

たことがある。しかし、このとき、内容の目新しさよりも、どちらかというとネーミングの上手さのほうが先行した印象をもっている。今回も、当初は「ビッグデータ」や「データサイエンス」が、統計学にドラスティックな変化をもたらすとは思えなかった。少なくとも筆者はそうであった。ところが、二〇一三年、西内啓氏による『統計学が最強の学問である』がダイヤモンド社から出版されて以来、統計学への追い風が社会現象として顕在化した。この本で紹介されたのが、当時Googleチーフエコノミストのハル・バリアン氏による「私はこれからの一〇年で最もセクシー（魅力的）な職業は統計家だろうって言い続けてるんだ（括弧内は筆者が加筆）[1]」である。これが二〇〇九年のことである。まさに、統計学にとって追い風である。一〇年たった今、統計学を取り巻く環境はどのように変化したのであろうか。

満を持して、文部科学省が動いた。二〇一二年度から五年間にわたる事業「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」である。イノベーションの推進を念頭に、新たな課題を自ら発見し、データに基づく数量的な思考による課題解決の能力を有する人材の育成をねらったものである。事業を採択した九大学（代表校：青山学院大学、東京大学、大阪大学、総合研究大学院大学、多摩大学、立教大学、早稲田大学、同志社大学、滋賀大学）は、統計関連学会（日本統計学会をはじめとする六学会）の支援を得て、コンソーシアムを

組織し、「統計教育連携ネットワーク」を設立し、現在に至っている。二〇一一年から始まった「統計検定」は、関連組織である統計質保証協会によるものである。

二〇一四年、日本学術会議情報学委員会による「ビッグデータ時代に対応する人材の育成」と題した報告書には、データ中心科学を専門とする教育組織の設置や基幹的研究組織内における恒久的なデータ解析部門の設置などの提言があった。二〇一五年六月の閣議決定「科学技術イノベーション総合戦略二〇一五」では、欧米と比較し、データ分析をスキルとする人材や統計科学を専攻する人材の不足が指摘された。これまで、我が国の大学には、統計学を専門領域とする学部／学科がなかったのである。これら統計学への追い風が、まず、滋賀大学の改革を生んだ。

3　統計教育の改革

二〇一七年度、滋賀大学に我が国で初めて、統計学を専門領域とした学部（データサイエンス学部：定員一〇〇名）が設立された。前学長の佐和隆光氏のリーダーシップのもと、東京大学から竹村彰通氏を招き、創設した学部である。当時、国立大学における文系不要論が

囁かれていた。経済学部と教育学部しかもたない滋賀大学では、ある意味生き残りをかけた改革の議論があったと思われる。統計学への追い風を背景に、文理融合型学部構想へと舵を切り、データサイエンス学部の設立に至った。まさに起死回生の舵取りであったに違いない。

学部紹介のホームページ[2]には次のような紹介がある。

「データサイエンスとは、社会に溢れているデータから《価値》を引き出す学問です。ICT（情報通信技術）の進化した現代では、あらゆるビジネスや医療、教育、行政等においても、高度なデータ処理能力、データ分析力が必要となっています。データから有益な《価値》を引き出すためには、これらの能力に加え、様々な分析経験を積むことが求められています。」

翌二〇一八年度には、岩崎学氏（当時、日本統計学会会長）を招き、横浜市立大学にもデータサイエンス学部が創設された。ホームページ[3]には、「ただの数字で終わらせるか。世界を変える力を引き出すか。」という刺激的なキャッチフレーズがある。

海外では、もともと統計学分野の独立した学部・学科が存在する。米国では Department of Statistics あるいは、Department of Biostatistics などの学科、専攻が数多くある。独立した組織があると、他領域（情報、数学、経済）と教育プログラムを構成しやすい。前述したように、

我が国には統計関連の学科がなく、分野点在方式であった。既存部局間の複雑な事情をもつ総合大学では、設置が難しい。しかし、滋賀大学は学長がリーダーシップを発揮しやすい環境にあった[4]。

滋賀大学の改革は、新学部の設立だけに留まっていない。二〇一六年、文部科学省は「数理及びデータサイエンスに係る教育強化」の拠点校として、北海道大学、東京大学、京都大学、大阪大学、九州大学の旧帝大に加え、滋賀大学を選定した。選定要件の一つとして、センターの設置が謳われている。滋賀大学では二〇一六年四月に、データサイエンス教育研究拠点として「データサイエンス教育研究センター」を設立し、先端的な教育研究活動を行うとともに、企業や自治体との連携、多様な大学間連携を通じて、様々な分野における新たな価値創造、社会貢献、教育開発を行っている。特に、産業界との連携には、同学経済学部の同窓会組織である陵水会の存在が大きい[5]。彦根高商時代からの伝統の重みを感じる。なかでも、トヨタグループとの連携によるビッグデータ分析人材育成プログラム「機械学習実践道場」は産学連携事業として注目に値する。

これら二つの学科紹介文から、データサイエンスとは「データから価値を引き出す科学」と言ってよい。「テクノロジー」ではなく「サイエンス」である。科学を「対象の本質的な

性質を探求すること」であるとするならば、データサイエンスは、その「探求」において、観察や実験によって得たデータを統計的に解析し、その結果を「対象の性質の探求」のために解釈する一連の統計的思考であると言える。これは自然科学領域でも社会科学領域でも不可欠な研究過程である。やはり「サイエンス」なのである。すなわち、データサイエンスは単なるデータ解析ではなく、価値創造へつながる科学的探求領域の学問である[6]。

「統計はもともと、「科学の文法（Grammar of Science）」として体系化された方法論で（Pearson 1892）、自然、社会、経済、人間行動のあらゆる研究課題に対して、データに基づく科学的探求のプロセスを提供し、（中略）、今日の「計量」を冠する多くの研究領域の基本ツールとなってきた[6]。」「データに基づき、科学を創造するプロセスを提供する科学」とする発想が当初から存在したのである。ただし、滋賀大、横浜市立大のホームページにあるように、データサイエンスの目的は価値創造にある。統計学は、ビッグデータという追い風に乗って、目的を価値創造とし、そのプロセスをデータに基づく科学的探求とした「データサイエンス」に衣替えしたのである。ただし、ビッグデータという追い風を駆動エネルギにするために情報学との融合を必要とした。すなわち、データサイエンスは、統計学、情報学、価値創造の三つの要素から成る科学である[5]。

統計学への追い風は大学教育にとどまらず、初等中等教育にも及んでいる。二〇〇五年、日本統計学会、日本品質管理学会をはじめとする一七の関連学協会が、「21世紀の知識創造社会に向けた統計教育推進への要望書」を文部科学省に提出した。その後、小学校と中学校の二〇〇八年告示の新学習指導要領において、統計内容の大幅な拡充と必修化が実施されている[6]。

例えば、我々の世代の小学生時代、グラフの読み方は社会科で登場し、算数ではなかったと記憶している。しかし、二〇〇八年の小学校・中学校の新学習指導要領では、図、表、グラフの教育が、社会科や理科に加え、小学校一年生から算数でも位置づけられている。中学校では数学で確率・統計の領域が設置され、二〇〇九年の高等学校の新学習指導要領では、数Ⅰで「データの分析」の単元が、また、数学Bに「確率分布と統計的な推測」の単元が挙げられている。

また、特筆すべきは、データに基づく科学的な問題解決力をコンピテンシーとして定着させることが新学習指導要領に謳われていることである[6]。これはまさに我が国の品質管理分野において体系化されてきた統計的品質管理（Statistical Quality Control：SQC）であり、「データに基づいた問題発見と問題解決」の方法論である。すなわち、我が国の製造業が得意

としてきた「改善」のプロセスにおける方法論が、初等中等教育に反映されたのである。

4　データサイエンティストに関わる不易流行

食品スーパーの顧客の購買行動分析を依頼されたことがある[7]。解析対象期間は二〇一二年一〇月一日から二〇一三年五月三一日の八ヶ月二四二日間である。この期間の売り上げデータとそれに対応するポイントカードのデータを入手した。ポイントカードシステムによって、買上二〇〇円ごとに一ポイントが付与され、あるインセンティブが与えられる。ＩＤ－ＰＯＳデータとは、商品販売時点データであるＰＯＳデータに顧客情報が付加されたものであり、「どんな人が」「何を」「何と一緒に」「何の購入の前に（後に）」購入したかという顧客個人における販売履歴情報である。

企業との共同研究の場合、頂いたデータの加工に多くのエネルギを費やすケースが多い。本事例の場合、入手したＰＯＳデータはレシートをテキスト化したテキスト形式データであった。まずこれをｃｓｖ形式へ変換する作業が必要であり、さらに、これにポイントカードシステムの個人情報をリンクさせることでＩＤ－ＰＯＳデータを作成した。この作業を面

倒にさせたのは、レジの売り上げ入力とポイントカード入力のタイムラグである。レジの売り上げ入力とポイントカード入力は別の機器で行われていたためである。しかも、ポイントカード入力時の金額は必ずしもレジの売上額と一致しない。二〇〇円ごとに一ポイントであることから、しばしば一〇円以下の桁が端折って入力されてしまう。つまり、ＩＤデータとＰＯＳデータのマッチング精度を上げるのに戸惑った。ここまでの処理を行って、初めてＩＤ－ＰＯＳデータとなる。結局、本学情報工学科の伊藤孝行研究室に助けを求めることになった。

この事例のデータを一般的なビッグデータとよぶかどうかは分からないが、とにかく、データ解析をする以前に、データ加工という大変な仕事が必要であった。情報技術に明るくなければ対応できない。このように、世の中の素データは、そのままＥＸＣＥＬシートに収まるような構造化データで存在することは稀である。ＳＮＳのデータ、画像データ、音声データなどは、その典型である。非構造化データを構造化データに加工するには、情報技術が必要となる。

構造化されたデータになって、はじめてデータを解析の俎上に載せることができる。次のステージは統計学の出番である。データ解析とその結果を解釈して、そこから何らかの価値

を生み出すことになる。上記の事例では、事例の対象である食品スーパーとの議論、および、過疎地問題から、「買い物弱者」への具体策として、移動スーパーの品揃えに対する情報提供という価値を議論した。価値創造のステージでは社会学の出番である。

繰り返すと、上記事例は、レジのデータとポイントカードのデータの入手に始まり、買い物弱者への対策という価値に至る過程で、情報学、統計学、そして、社会学の知識を必要とした。このように、統計学、情報学、価値創造の三つの分野の素養を併せもった人材がデータサイエンティストである。滋賀大学のデータサイエンス学科のカリキュラムも、データエンジニアリング系（情報関連）科目、データアナリシス系（統計系）科目と価値創造科目（経済、経営系科目、多分野における価値創造の実例紹介、価値創造の実践等）から成る。データエンジニアリングとデータアナリシスの能力をベースとして、価値創造に挑む、「逆Ⅱ型人材[5]」の育成をめざしている。しかし、学部教育で三つの分野の素養を併せもつ人材育成となるとなかなか難しい。例えば、統計学のスペシャリストをめざし、リテラシーとして情報学を学び、ＰＢＬ教育を通じて、「統計学に裏打ちされたコミュニケーション能力」を涵養する、と言った人材教育が想定できる。あるいは、ある専門分野をもち、イノベーション志向をもつ社会人が「逆Ⅱ型」のベースとなる統計学と情報学のスキルアップをめざす、と

いった人材教育も想定できる。

今まで統計学は黒子であった感が強い。特に、工学の分野ではその傾向が強い。例えば、生産性を上げるべく新工法が提案されたとき、その報告なり、論文の主役は新工法の基礎となった要素技術である。新工法の発見や検証に実験計画法が使われ、効率よく新工法の提案ができたとしても、データ解析である実験計画法の部分は「黒子」である。実験計画法は要素技術（主役）に対する管理技術（黒子）である。医学、薬学の分野ではデータおよびデータ解析がエビデンスとして要求されるが、それでもこれまではデータ解析が主役であったとは言えない。ビッグデータの追い風に乗って、「黒子」であった統計学が「データサイエンティスト」という名の下で主役に抜擢された感がある。無論、前述したように、情報学という共演者の存在があっての主役であり、また、価値創造へのプロセスは「不易」であるが、対象とするドメインにおけるデータの質／量の多様性への新たな対応が必要となる。

5　データおよびデータ解析の不易流行

ビッグデータとは何か？当初は、「Volume（量）、Variety（多様性）、Velocity（速度）のい

ずれも大きい性質を備えたデータ」と説明されていた。日本学術会議「ビッグデータ時代に対応する人材育成」の提言[8]によると、「市販されているデータベース管理ツールや従来のデータ処理アプリケーションで処理することが困難なほど巨大で複雑なデータ集合の集積物を表す用語」と定義されている。「巨大」はVolumeに、「複雑」はVarietyに対応すると思われるが、Velocityの対応は見当たらない。Velocityを「更新速度」と解釈することもあるが、データ獲得の速度がそれほど一般的であるとは思えない。どうもビッグデータに対する学問的な定義はできそうにない。しかし、「今はビッグデータ時代です。」と言えば、それは理解できる。スマートフォンの普及を考えれば、明らかに、巨大で複雑なデータが安価に入手できる時代になったことは実感できる。そこで、本稿では「ビッグデータの時代のデータ」をとりあえず「ビッグデータ」とよぶことにする。前述した食品スーパーの素データであるPOSデータは、レシートをテキスト化したテキスト形式データであった。このデータはサンプル×変数の構造化データではない。テキストデータや画像データは非構造化データであり、「複雑さ」という意味で言うならば、このようなデータがビッグデータの例と考えてよい。

「巨大」で「複雑」なデータであるビッグデータは、これまでのデータとは何が違うのだろうか？この問に対して、従来の統計学と異なる点を強調したいがためか、誤解ではないか

と思われる記述が散見される。例えば、「従来の統計学は母集団からサンプリングされた標本データを解析対象とするものであり、これに対してビッグデータはデータが巨大であることから、データが母集団そのものである」というものである。データが解析対象である集団に対して、サンプリングされたものか、全数かどうかは、母集団の大きさとサンプルの大きさとの関係であり、サンプル数自体とは別の話である。ビッグデータの解析目的の中心は予測・判別である。これらは、新たな標本に対する予測・判別である。ビッグデータの場合も対象とするのは、データの背後にある無限母集団である。さらには、「従来の統計学は仮説検証の演繹型であり、一方、ビッグデータの解析は帰納型である」という誤解である。統計学が仮説検証の演繹型か仮説生成の帰納型かは、解析目的によるのであって、サンプル数の大きさとは別の話である。違いを強調したいがための誤った記述は普及に悪影響を及ぼす。

ビッグデータは明らかに、統計学の発展を促した。以後は、ビッグデータの登場に至るこれまでの統計学の不易流行について記述する。統計学は、前述した「黒子」的な役割の時代から、今日の「主役」に躍り出るビッグデータの時代に至るまで、脈々と発展・普及してきた歴史をもつ。

まず、統計学の発展に寄与した歴史上の人物を挙げるならば、F．ゴルトン（一八二二

～一九一一）、K．ピアソン（一八五七～一九三六）、R．A．フィッシャー（一八九〇～一九六二）、J．ネイマン（一八九四～一九八一）であろう。ゴルトンは親と子の身長を関係づけるために散布図を用い、相関係数、回帰直線を考案した。それを理論づけしたのがピアソンである。フィッシャーは実験計画法を体系化し、標本分布（例えば、正規分布の標本分布の一つであるF分布）を導出した。ピアソンとフィッシャーの二人が近代統計学の礎を築いたといえる。ただし、二人の統計学に対する立場は異なる。ピアソンの統計学は記述統計であり、一方、フィッシャーは推測統計の立場である。推測統計の立場をより鮮明にしたのがネイマンである。ネイマン－ピアソン（E．S．ピアソン：K．ピアソンの息子）の基本定理で有名なネイマンは、この基本定理をもとに、仮説検定における帰無仮説と対立仮説、有意水準と検出力などを考案した。現在、ほとんどの統計学の教科書は、ネイマン－ピアソン流で書かれている。著者もネイマン－ピアソン流で書かれた教科書で統計学を学んだ一人である。この流れが、前述した、統計学＝推測統計、あるいは統計学＝仮説検定の誤解を生んだと考えられる。

記述統計とは「データの特徴を記述することを目的としたもの」であり、推測統計とは「データの背後にある母集団を特徴付ける分布の母数や母集団の傾向をデータから推測

することを目的としたもの」である。ピアソンの研究の背景にはゴルトンの遺伝学があり、フィッシャーの研究の背景には、自らが体系化した農業における実験計画がある。歴史的な推移は、記述統計から推測統計であるが、推測統計が記述統計に取って代わったというわけではない。統計学が対象としたドメインが、遺伝学から農業に拡大したのである。また、記述統計に推測統計の要素が欠けていたわけではない[9]。例えば、カイ二乗適合度検定はピアソンが考案したものである。これは「想定された理論分布がデータを記述するものとしてデータと十分整合するか否かを見る[9]」ものであり、記述統計の立場である。フィシャーにとって、検定とは差（肥料の効果）があることを立証する意味があったのに対し、ピアソンにとって、検定とは（遺伝学の）法則の成立を創造する意味があった。すなわち、ピアソンにとって、統計的法則はデータを記述するための枠組みであって、データの中に存在するものではない。「データの背後に法則があると考えるのではなく、法則はデータの挙動よって創造される」というものである。この考え方は、ビッグデータの解析の特徴である「データ駆動型の解析」につながるものである。統計学の不易流行を考えたとき、この点は注目すべきである。ピアソンの思想はビッグデータの時代に至るまで「不易」なものである。

統計学が対象とするドメインは遺伝学から農業へ、そして工業へと拡大した。具体的な応

用は品質管理における抜取検査である。ネイマン・ピアソン流の仮説検定は、ロットが「合格か不合格か」を意思決定する抜取検査の理論にマッチする。ネイマン・ピアソン流の統計学は、まさに統計的意思決定である。ちなみに、農業をドメインとしたフィッシャーの実験計画を、工業（特に、設計）のドメインに拡大したのが田口玄一である。

フィッシャー流とネイマン・ピアソン流の違いは以下の通りである。データ解析ソフトを使って仮説検定（例えば、二つの母平均の差の検定）を行ったとき、P値が出力される。P値とは「帰無仮説が成立している下で、算出した検定統計量の値以上となる確率」である。ネイマン・ピアソン流であれば、前もって決めた確率（五%がよく使われる）よりP値が小さいとき、「差がない」という仮説（帰無仮説）を棄却し、「差がある」という対立仮説を採択する。前もって決めた確率を有意水準という。P値と有意水準の大きさを比べるだけであり、P値の大きさは問わない。この結論の解釈は「私は同じ検定を一〇〇回行ったとすれば、平均して五回は間違いを犯しますが、帰無仮説を棄て、対立仮説を採ります」である。フィッシャーはこれに異を唱えた。「抜取検査であれば、母集団がいくつもあり、検定を繰り返すことが想定できる。しかし、自然科学における母集団は仮説的なものであり、検定を一〇〇回行うことは想定できない。」と。ネイマンは「確率＝頻度」論者であり、一方、フィッ

シャーは「確率＝確からしさの尺度」とする対立である。P値を考案したフィッシャーは、P値をエビデンス力と解釈した[10]。すなわち、フィッシャー流では「P値が小さいと「差がある」というエビデンス力が大きい」と解釈する。

我々はしばしば、有意水準一％で有意なときは「高度に有意」として「＊＊」をつけ、有意水準五％で有意なときには「有意」として「＊」をつける。これを、「有意水準一％で有意」のほうが、有意水準五％に比べて「差があることが確からしい」と解釈する。この解釈は、ネイマン‐ピアソン流とフィシャー流の折衷である。

ここで、P値に関する誤解をあげておく。これらは、P値はサンプル数に依存するという認識がないことから生じる誤解である。

1）P値が小さいと、より効果が大きいと判断する
2）サンプル数を無視して、P値で判定する
3）サンプル数の異なる検定のP値の比較をする

上記2）の誤解は、ビッグデータとも関連する。ビッグデータはサンプル数が「巨大」である。サンプル数が大きいとP値は小さくなる。つまり、極々小さな差であっても「有意な差」と結論づけてしまう。サンプル数が「巨大」であるビッグデータに仮説検定は馴染まな

い。また、統計的に有意差があることと、対象のドメインにおいて差があることとは異なる。仮説検定はアクションへの動機付けであり、具体的なアクションは推定に依るべきである。

ここまで、統計学は応用のドメインを遺伝学→農業→工業へと拡大していった。この拡大をサンプル数でなぞったならば、大まかに、大（遺伝学）→小（農業）→中（工業）であると言える。そして次の時代が、ＳＮＳで代表されるように、サンプル数が「巨大」である「社会」というドメインである。

ビッグデータの時代において、注目すべきはベイズ統計の復活である。二〇〇七年朝日新聞（別紙『be』）に「三〇〇年後に脚光ベイズの定理」が掲載され、人工知能への応用が注目された。「復活」と表現したのは、これまで統計学の一般の教科書において、ベイズの定理の記述はあってもベイズ統計の記述はほとんどないと言ってよいからである。これは、ピアソン、フィッシャー、ネイマンがともに、ベイズ統計学を否定したことが大きいと思われる。頻度論者であるネイマンが母数の事前分布を否定したことは容易に理解できる。ただし、フィッシャーは、母数を変数とした関数である「尤度」の概念を発展させたことから、また、母数の事前分布を考えるベイズ統計を、現象を説明する一つの考え方であるとすれば、ピアソンもネイマンほどベイズ統計を完全否定したとは思えない。

ベイズ統計は、解析対象とする統計モデルを特徴づけるパラメータの事前分布を想定するものである。いわゆる、解析時での事前情報である。事前分布からデータを得ることによって事後分布を得るという考え方である。データから母数の事後分布を得るというアプローチは、まさにデータ駆動型の統計であり、データ、しかもビッグデータによって事前分布から事後分布を「学習」するという発想が人工知能に応用されている。

コンピュータの処理能力の著しい向上とビッグデータの存在は、情報学において機械学習の発展をみた。代表的な手法が、ニューラルネットやサポートベクトルマシンである。ただし、ビッグデータと同様に、機械学習にきちんとした学問的な定義があるわけではない[11]。機械学習の特徴としては、非線形モデルであること、用途が未知データに対する予測に重点が置かれていることである。予測への用途は、モデルに汎化能力が要求されることになる。モデルが非線形であることから、学習したモデルのパラメータ学習のロバスト性、あるいは、モデル構築に用いたデータに過度にフィットする過学習が問題となる。その対策として、スパース推定などの研究が進められている。

予測に対する機械学習の成果は目を見張るものがある。しかし、予測モデルは必ずしも因果モデルではないことに注意が必要である。電気炉の内部温度を外壁に取り付けた温度セン

サーの値から予測するモデルでは、予測の方向は、外壁から炉内部の方向である。一方、発熱体は炉内部にあるので、因果の方向は炉内部から外壁の方向である[12]。予測に因果関係は必ずしも必要ない。予測モデルのパラメータの値を解釈する必要もない。

また、相関関係は必ずしも因果関係を表さない。四七都道府県の年間の犯罪発生件数とコンビニエンスストアの店舗数には高い正の相関がある。だからといって、「コンビニエンスストアは社会悪である」とは誰も言わない[13]。国別のチョコレートの消費量とノーベル賞受賞者数は高い正の相関関係がある。だからといって、チョコレートが知能を高めるわけではない（例えば、[14]）。逆に、相関関係はないが因果関係があるというケースもある（例えば、[13]）。ところが、相関関係の存在を知ることだけでは不十分であり、因果モデルからアクションの効果を知りたいケースが多くある。いわゆる介入効果を知りたいのである。たとえビッグデータといえども因果モデルの構築は簡単ではない。ビッグデータは典型的な観察データである。観察データでは共変量の存在が介入効果の推定の邪魔をする。効果が大きい共変量が観察されていないと因果分析ではミスリーディングを招く。

ところが、共変量の存在をものともしないで、少数データから介入効果を推定する方法がある。実験である。フィッシャーの出番である。ビッグデータが観察データであるのに対し

て、実験によるデータを計画データとよぶ。実験は因子の水準を意図的に変更するという介入によって、介入効果を推定することができる。これが実験の最大のメリットである。共変量の影響は実験誤差として残るが、共変量の効果が介入効果に含まれることはない。観察データが主流である社会科学の分野では実験を計画することは難しい。そこで、仮想的に実験の場を作る工夫が提案されている（例えば、[15]）。

予測や分類以外にサンプルが「巨大」なビッグデータであることを利用した解析の方向性がある。その方向性を、サンプル数が「巨大」であることを利用した統計的因果探索[16]にみることができる。この方法は、ビッグデータから分布を求めることを基本としたものである。我々は、中心極限定理のおかげで、フィッシャーの標本分布論に始まる正規分布を仮定した推測統計の美しい理論体系を幅広く利用できる。正規分布を仮定した時点で、平均（一次モーメント）とばらつき（二次モーメント）を求めれば十分である。正規分布は平均とばらつきによって決まる大変便利な分布だからである。ところが、このことは、正規分布を仮定した時点で三次モーメント以上の情報は捨てていることを意味する。もっとも、少数データでは三次モーメント以上を求めたところで推定精度が著しく悪く、使い物にならない。ところがビッグデータであれば正規分布を仮定せず、分布を拠り所とする解析、すなわち、三次

モーメント以上の情報を解析対象とすることができる。これこそビッグデータのなせるわざである。この発想はいろんな分野で応用可能である。例えば異常検知にも活用できる。個々のデータ、あるいは、平均値やばらつきをモニタリングするばかりではなく、分布系をモニタリングすることによって、異常の兆候を早期に検知することが考えられる[17]。

ビッグデータであれ少数データの計画データであれ、データサイエンスの目的である価値創造に対してそれぞれの役割をもっている。「統計学よ、さようなら。機械学習よ、こんにちは。」でもなければ、その逆でもない。

6　おわりに

統計学は、遺伝学→農業→工業→社会と、そのドメインを拡大しながら発展をみた。むろん、ドメインは医学、薬学あるいは心理学なども含む。今や、統計学の目的はデータの獲得からデータ解析による対象ドメインでの価値創造であると言ってよい。そのプロセスは、ピアソンの「文法の科学」を根源にもつ不易流行の統計的科学思想であり、まさにデータサイエンスである。

統計学の不易流行を語るときに、前掲ゴルトンのいとこであるF．ナイチンゲール（一八二〇～一九一〇）は外せない。一八五四年に英国がクリミア戦争に参戦した際、従軍看護婦として戦地に赴いた話は有名である。そこで、彼女はエビデンスとして、データをビジュアル化した「鶏のとさか」グラフを使い、病院内の衛生管理の重要性を訴えた。彼女は統計学者でもある。ナイチンゲールのことばを引用する。「神の御心を知るには統計学を学ばなければいけない[18]」。冒頭のGoogleチーフエコノミスト、ハル・バリアン氏のことばはナイチンゲールから約一五〇年後である。

参考文献

1　西内啓（二〇一三）『統計学が最強の学問である』ダイヤモンド社
2　滋賀大学データサイエンス学部：https://www.ds.shiga-u.ac.jp/about/ds/introduction/.
3　横浜市立大学データサイエンス学部：https://www.yokohama-cu.ac.jp/academics/ds/index.html.
4　竹村彰通、佐和隆光、吉川英治、姫野哲人（二〇一五）「データサイエンスの大学専門教育」　日本品質管理学会第四五回年次大会
5　竹村彰通（二〇一九）『データサイエンス入門』岩波新書，岩波書店
6　渡辺美智子（二〇一三）「知識基盤社会における統計教育の新しい枠組み」日本統計学会誌，Vol. 42, No. 2, 253-271.
7　Terakado, M., Nishina, K. and Ishii, N.(2015):"A Case Study on a Measure against Food Deserts Issue",The

3rd Asian SME Conference 2015.
8 日本学術会議情報学委員会（二〇一四）「ビッグデータ時代に対応する人材育成」
9 竹内啓（二〇一八）『歴史と統計学』日本経済新聞出版社
10 柳川堯（二〇一八）『P値～その正しい理解と適用～』近代科学社
11 竹村彰通、姫野哲人、高田聖治編（二〇一九）『データサイエンス入門』学術図書出版社
12 椿広計（一九九四）「回帰分析から因果分析へ」標準化と品質管理、Vol. 47,No. 5, 111-116.
13 仁科健、川村大伸、石井成（二〇一八）『スタンダード品質管理』培風館
14 黒木学（二〇一七）『構造的因果モデルの基礎』共立出版
15 伊藤公一朗（二〇一七）『データ分析の力 因果関係に迫る思考法』光文社新書、光文社
16 清水昌平（二〇一七）『統計的因果探索』講談社
17 吉野睦（二〇一七）「適応的制御時代の工程管理」統計関連学会連合大会
18 渡辺美智子（二〇一三）「身近にある統計［改訂版］」品質月刊テキストNo. 396, 品質月間委員会

Immutability in Developing of Statistics

These days, big data and data science are full of media. That is exactly a tailwind to statistics. In 2017, the Faculty of Data Science was established at Shiga University, the first department in Japan that specializes in statistics. The purpose of data science is value creation in target domain through data acquisition and analysis. In this paper, we review the history from modern statistics beginning with K. Pearson to data science in the big data era, and discuss continuity and change concerning statistical education, data scientists, and data analysis. We can recognize that while expanding the domains with genetics (F. Galton and K. Pearson), agriculture (R. A. Fisher), industry (J. Neyman) and society (the big data era), statistics has been developing in responses to the diversity of data in these domains. Behind that lies the immutable scientific thought from Grammar of Science by K. Pearson.

仁科健 | Ken NISHINA
愛知工業大学・名古屋工業大学
品質管理・応用統計学
経営学部教授・名誉教授