データサイエンス今昔

横幹連合理事　田村 義保*

1. はじめに

2019 年 2 月の横幹理事会終了時に原稿執筆の依頼を受けました．某学会の会報の巻頭随筆を書き上げた直後であったので，自信たっぷりに気軽に引き受けてしまいました．最近の巻頭言を読ませていただきましたが，学術界，産業界や政策の昨今の状況に対する哲学的な分析が行われていました．深い哲学的な思考をほとんどしない著者（いくらなんでも，少しはするので,「深い」という修飾語をつけています.）としては,「うーん，難しいな」と思ってしまっています．しかし，何かを書かなければならないという「脅迫観念」もあり，責任感だけは豊富なので，執筆を開始しました.「単なるボヤキ」(「ボヤキ」でも Twitter でやれば立派な意見？）にはならないように注意して書いています．

「あなたの専門は何ですか？」という問いに対しては，計算統計学とか非線形・非平衡の統計物理と答えています．世間では物理乱数の研究者，スパコン大好き研究者と思われているようです．広いが，浅くいろいろなことをやってきたと思っています．

しかし，本稿では，統計教育，統計不正，EBPM, Society 5.0，データ中心科学等の最近注目されている事柄に焦点を当て，最近，少しだけ深く考えていることについて述べたいと思います．

2. 最近思うこと

初中等教育だけでなく高等教育や社会人教育にいたるまで，統計教育，データサイエンス教育が注目を浴び，盛んに実施されています．現在の学習指導要領は小学校では平成 23 年度，中学校では平成 24 年度から全面実施されました．高等学校では数学・理科は平成 24 年度から，他は平成 25 年度から全面実施されています．次の学習指導要領もすでに発表されており，後，2 年もすれば小学校から順次実施されて行きます．現課程よりも内容は充実しており，日本人の平均的な統計リテラシーの底上げが期待されます．高校においては，数学科だけでなく情報科でも「情報活用能力」を身に付けるために,「情報 I」という科目が新設されています．教員用研修教材が一時期，文科省 Web で公開されていましたが，2019 年 6 月 1 日時点ではなぜか準備中に戻っていました．少しだけ読ませていただきましたが，良く書けてはいましたが，内容が膨大で,「この内の何割を教えるのだろうか？」,「いったい週に何コマあるのだろうか？」,「高校の先生は大変だな.」と思ってしまいました．2018 年 3 月には「情報科」の関係者がプログラミング言語は決めていないと言われていましたが，少なくともデータ解析については R や Python を想定しているということを聞いています．私はデータサイエンスリテラシーの教育を主目的とし，プログラミング能力向上を目指さない単元の指導においては，GUI 等を用いて，正しく解析でき，結果を正しく理解・分析できるようにするための教育に徹するべきであると考えます．もちろん，プログラミングの基礎教育も重要です．

総務省政策統括官（統計基準担当）による,「生徒のための統計活用～基礎編～」,「高校からの統計・データサイエンス活用～上級編～」も刊行されており，初中等教育レベルにおけるデータサイエンスリテラシー教育は充実して行われていると考えます．関係者のご努力に敬意を表したいと考えます．

*統計数理研究所　統計思考院　東京都立川市緑町 10–3

Received: 12 June 2019.

大学・大学院でのデータサイエンス教育は，“scienceandtechnology.jp/archives/13026”（データサイエンスが学べる日本の大学・学部　～データサイエンティストになる方法～）（2019 年 6 月 1 日アクセス）によれば，「データサイエンスが学べる学部，学科，コースが最近設置された（される予定の）大学」が 21，「データサイエンスが学べる（可能性のある）その他の大学，学部，学科」が 24 もありました．滋賀大学や横浜市立大学にデータサイエンス学部が設置され，2019 年度から滋賀大学に 2020 年度から横浜市大に大学院が設置され，高度な知識，データ解析能力を有した専門家を育成する体制が整いつつあることは喜ばしい限りです．

また，文部科学省の数理・データサイエンス教育強化拠点コンソーシアム（北大，東大，滋賀大，京大，阪大，九大）事業においては，数理・データサイエンスを中心とした全学的・組織的な教育を行うセンターの整備と全国的なモデルとなる標準カリキュラム・教材の作成等，高等教育の充実に取り組んでいます．統数研は総合研究大学院大学統計科学専攻で後継研究者を育成するだけでなく実務家として高度なデータ解析能力を有するデータサイエンティストの育成にも努めています．

このような教育体制が 20 年以上前から整っていたら，現在，問題になっている「統計不正」は起こらなかったでしょうか？その答えは「是」でもあり「否」でもあります．また，多くの新聞報道であった，日本の統計職員の人数の少なさが原因なのでしょうか？著者が参考にした産経新聞の Web 記事（2019 年 2 月 9 日 18 時 57 分）によれば，人口 10 万人あたり，一番多いのはカナダの 13.9 人，次が英国の 10.5 人であり，日本はわずか 2.1 人しかいないことになっていました．ドイツの 2.0 人よりは多いですが，日本の統計制度のもう一つの問題点は，省庁ごとの「分散型」にあると考えます．高等教育機関での統計研究・教育も分散型であることが関係しているのかもしれませんが，経験の蓄積，継承という意味では非効率的な組織と言えます．

また，かなり「ひがみ根性的な」意見となりますが，「統計なんて，優しいから誰でもできる」という考えが蔓延していて，統計関係業務の担当者に，それほど知識が無い方を定期的な異動により従事させてきたのかなという疑いも持っています．はっきり言って，基幹統計の調査目的決定，調査方法選定，とりまとめ，分析には高度の知識が必要であり，かなりの経験が必要なように思います．最近，いろいろな基幹統計の担当部署の関係者の話しを聞く機会がありますが，ほぼすべての方は，統計的な知識も十分にあり，真摯に業務を行っているように思います．したがって，上述の「ひがみ根性的な」発言は間違いであるのかもしれません．では，毎月勤労統計調査等の厚労省の統計不正はなぜ起こったのでしょうか？「魔がさした」のでしょうか？これは，現政権に忖度して，数値補正を行ったのではなく，行う方が正しいので，やったにすぎないのです．それにもかかわらず，一部のマスコミによる「忖度のために数値を修正した」という報道は，全く統計を理解していないのか，理解しているが，このような記事の書き方の方が国民に受けるから書いているのかよく分からないところです．この補正により，統計不正が明らかになったと言えます．この問題については，「統計 2019 年 6 月号」（一般財団法人・日本統計協会）を是非とも一読して欲しいと考えます．

補正を行うことも統計委員会に届ける必要はあったと思います．しかし，全数調査の結果に乗率の補正は要りませんから，届け出ることは不可能です．サンプリング調査への変更を計画した時点で，統計委員会の許可が必要でした．長い間，放置されてきた理由は良く分かりません．最近になっても統計不正に関する報道は続いており，2019 年 5 月 28 日の読売新聞には自民党による提言が紹介されていました．「法人調査のオンライン化」や「大学での実務教育」というのが含まれているそうです．オンライン化は小規模事業所の負担が増えるというような観点から実施されて来なかった可能性もありますが，著者がある調査のオンライン化のワーキンググループに出たのは 21 世紀になってすぐです．いまだに言っているのと思ってしまいます．実務養育はもちろん必要です．データ解析の教育はすでに述べたように進んで行くと思いますが，データを作成するための方法である調査方法や実験計画等の教育はそれほど重視されていないように思

います．せっかく，データサイエンス教育を行うのであれば，データを取るところからやって欲しいものです．臨床試験や超高価な測定装置を利用するビッグサイエンスでの実験・観測によるデータ取得もより精緻にやって欲しいものです．

では，公的データや臨床試験データを正しくとることができればそれだけでよいのでしょうか？EBPM や EBM という言葉が身近に満ち溢れているように感じています．とくに，EBM については，「EBM でない M は何か?」と考えてしまいます.「医は仁術」の世界を指しているのかもしれません．最近，高血圧の治療目標が改定されようとしていますが，どのような根拠に基づいているのだろうかと考えてしまいます．木曜日に発売される出版社系の週刊誌が，同じようなデータ（エビデンス）に基づいているにも関わらず，食品の安全性について真逆のようなことを言っているように思えます．所詮，週刊誌だからではなく，同一のデータに，異なってはいるが，正しい統計的データ解析手法を，正しく用いて，全く正反対の結論を出すことも可能です．

経済セミナー 2019 年 4・5 月号に，成田悠輔氏（イエール大学助教授）による『「エビデンスに基づく政策」に反対する』という論文があります．書きだしは，EBPM が良いということ自体がエビデンスに基づいておらず，意見に基づいている Opinion BPM であるとやや過激に書かれてはいますが，結びでは,『目的をデータ駆動で発見し，発見した目的のために，EBPM を使い』とあります．因果関係を測って，EBPM の良さを示したという例も紹介されています．EBPM を始める前に，方法の違いによる，結果の違いを推定するのは難しいかもしれませんが，統計的データ解析で注目されている因果推論や傾向スコアを用いるとできるかもしれません．

3. これから

前節で紹介した成田氏の論文にも「データ駆動」という用語が使われていました．北川会長による巻頭言にも「データ駆動」という用語が出てきます．北川会長の巻頭言では第 4 の科学は「データサイエンス」になっています．以前は「データ中心科学」と言われていたように思います．多くの人が想定している，高速計算機利用が重要な役割を持つ現時点の「データ中心科学」とは違っているとは思いますが，実験科学，理論科学，シミュレーション科学が生まれる前の，観測や観察結果から事実・法則を発見しようとしていた科学も「データ中心科学」であると言えるものと思います．エドモンド・ハレーやルネ・デカルトによる風向図は，風の可視化です．ヨハネス・ケプラーの 3 つの法則はティコ・ブラーエによる大量の惑星観測データから導かれたものと考えます．

ジョン・テューキー氏が提唱した EDA（Exploratory Data Analysis，探索的データ解析）は，ある意味では,「データ中心科学」を提案したものだと考えます．EDA はデータサイエンスの第一歩と位置づけられているそうですが，テューキー氏が提唱した時点で，数理統計学に流れ過ぎた統計コミュニティーを現実のデータ解析に引き戻したかは良くわかりません．しかし，上述のように現在のデータサイエンスでは重要な地位を占めていると思います．

ビッグデータがバズワードであるかどうかと議論されたのは「今は昔」という感じになっています．データマイニングとどう違うのかというのはいまだに議論されているかもしれません．しかし，重要なことは，現実にあるデータを活用して，科学的，社会的な結論を導き，科学の発展や生活の質の向上に寄与することです．

AI によるシンギュラリティがあり，無くなる職がいっぱいあるということを，耳にする機会は減りましたが，いまだに，AI 信仰は強いと思います．AI ターミナル（AI スピーカーではないでしょう）を使うと家事の一部が楽になるかもしれません．しかし，スイッチを入れるだけなら,「OK，＊＊＊＊」というよりも自分で入れた方が速いと思います．また，Society 5.0 時代を迎えてという書き方をする研究者もおられます．しかし，正しくは，Society 5.0 時代を目指してであると思われます．

統計不正が起きたことに関して，“https://www.itmedia.co.jp/news/articles/1903/13/news020_4.html”（“数字に弱いニッ

ポン”，統計不正で露わに　「データの歪み」なぜ放置？　国会議員×元日銀マンが斬る．2019 年 03 月 13 日 07 時 00 分 公開）（2019 年 6 月 1 日アクセス）があります．人材不足や予算不足ということが書かれています．著者が注目したのは，見出しの “数字に弱いニッポン”です．日本人は数字に無頓着すぎます．昨年，ものすごく速度の遅い台風の進行速度が時速 10km くらいになった時に歩くほどの速度になったと言い，時速 20km くらいになった時にジョギングくらいの速さになった言った超有名アナウンサーがいます．ジョギングで時速 20km の方なら来年のオリンピックでマラソンの金メダルがとれてしまいます．何気なく数値を見て，重要に思わない，日本人が多すぎるように思っています．私は，統計学の講義で学生に「数字に愛情を持とう.」と言っています．日本人に数学リテラシーが不足しているのは,「数字への愛情不足」が関係しているように思えてしかたがありません.

文科省が計画しているデータサイエンスに関する初中等教育，高等教育が成功して，教育を受けた方々のリテラシーが向上することを願います．マスコミ報道は必ずしも正しくないということに気づいて欲しいものです．データは正しいが，分析を疑えというようなことも言っています．一日も早く，公的統計データだけでなく，科学的な実験データ，企業による製品検査データの信頼を取り戻して欲しいと考えます．横幹に関係している言葉で言うといろいろなシステムの信頼性回復が必要です.

交通事故で，多くの方が被害を受けているので，話題に取り上げるのは気がひけています．アクセル，ブレーキの踏み間違いが事故原因のほとんどのように言われています．そうならば，踏み間違いに気づいてブレーキをかけるくらいのことはやって欲しいものです．今の車はデジタル制御なので，制御ミスでブレーキのつもりがアクセルになるようなエラーは現在でもほぼ 0 だとは思いますが，すべての操作でプログラムエラーを減らして欲しいものです.

「はじめに」に書いたことはほぼ書けたと思っています．私が好きなスパコンは「もの」だと思いますがモデルや数式等の「コト」を実現するためのものです．乱数の概念は「コト」であると思いますが，擬似乱数発生器も物理乱数発生機も「もの」です．統計数理研究所で 30 年以上も新しい物理乱数発生機開発に関係してきました．もっと速くて，優れた品質の乱数発生を目指して，今，しばらくは頑張りたいと考えています.

謝辞: 巻頭言を執筆させていただいことに御礼申し上げます.