特別論文

データサイエンスと統計的因果推論

横浜市立大学データサイエンス学部　岩崎　学

1. はじめに

AI、IoT、ビッグデータ、ディープラーニング、機械学習。これらの単語を目にしない日はなく、データサイエンスも日常語になりつつある。GoogleのチーフエコノミストHal Varianが「今後10年間で最もセクシーな職業は統計家である（I keep saying the sexy job in the next ten years will be statisticians.）」と言ったのが2009年で、Varianの言う10年は過ぎ去ってしまったが、統計家はデータサイエンティストに置き換わり、セクシーかどうかはともかくとしてその人気はとどまるところを知らないようである。言葉は置き換えられたが、統計学がデータサイエンスに置き換わったのではない。統計学がデータサイエンスの中核をなしていることは、識者でなくても論を俟たないであろう。もちろんデータサイエンスは統計学の別称では決してなく、それは、近年のコンピュータの高速化とネットワーク環境の普遍化およびストレージの大容量化をもたらした圧倒的な情報科学の発展に触発された新しい学問体系である。

本稿の目的は、そのようなデータサイエンスの全体像を俯瞰するものではなく、むしろその対極にあるともいえる統計的因果推論を基軸にして、データサイエンスの今後の確かな発展のために必要なことは何かを論ずることにある。そのため、第2節でまず一般的に研究あるいは日常業務の目的に応じた分類を述べ、それを達成するための方法論について言及する。第3節ではダイレクトマーケティングやWEBマーケティングで議論されることの多いA/Bテストを例にとり、何故それが有用であるかを統計的因果推論の見地から詳細に議論する。この節が本論の中心的部分である。続く第4節では因果推論の具体的な実行のために開発された最近のテクノロジーについて述べ、最後の第5節で簡単に全体をまとめると共に、さらなる統計的因果推論の理解に資するための文献の解説を行う。

2. 研究の分類とそのための方法論

データサイエンス的な研究あるいは業務では、とかくデータの量と質あるいは難解な分析ツールの扱いに翻弄され、本来の目的を見失ってしまう恐れがある。分析の目的およびそれを実現するための研究の種類を明確に意識しなくてはならない。ここでは、研究あるいは業務を目的別に、

［ｉ］現状把握、［ii］予測と判別、［iii］因果関係の確立

とし、方法論別に、

［a］実験研究、［b］観察研究、［c］調査

に分類する。これらの方法論のうち、実験研究と観察研究は共に、ある処置の効果の立証を目的とするという意味では類似であり、調査は必ずしもその目的を有していない。実験研究と観察研究の違いは、実験研究が研究対象の選択や処置の割り当てが研究者自らあるいはその指示の下に行われるのに対し、観察研究ではそれらが観察対象の判断に委ねられている点である。これらの分類を念頭に置き、最近の分析手法の特徴を概観しつつデータサイエンスの最近の動向を詳しく見ていく。

研究あるいは業務の第一歩は［i］の現状把握である。現状がどのようになっているかの正確な理解およびその背後に隠された構造の探索は、その後に続く分析にとってきわめて重要である。そこから研究課題が見えてくることも稀ではないであろう。この段階に資する近年のデータ可視化ツールの発展には目覚ましいものがある。特にビッグデータと称される超大量かつ日々刻々と変化するデータの集計やグラフによる可視化は、生のデータそのものを扱うことが不可能であるだけに必要不可欠なものである。それらビジュアルに優れたデータの提示、あるいはダイナミックな動画を駆使したプレゼンテーションツールの進展は特に著しい。そのようなデータの可視化により解決策が見出される課題・問題も少なくないであろうと推察される。［c］の調査がその方法論となる。

今や古典的と称される恐れもある多変量データ解析法の中の、主成分分析法、因子分析法、クラスター分析法などはこのカテゴリーの解析手法である。それらを包含する形で、現在のデータサイエンスツールの大部分がここに注力しているとも言える。機械学習の分野での教師なし学習と称される手法群がここに分類される。しかし反面、確かにグラフィックスはきれいで印象的であるが、だからどうなのかという疑問が呈されることも稀ではない。この段階にのみとどまっていて「次」がないのでは実際の研究課題の解決にならないという指摘も多く見られる。

「次」の段階である予測と判別［ii］は、実際問題上きわめて重要かつ有用でこれを目的とした方法論は数多い。予測は、何らかの特性値を既知の変量（説明変数）に基づいて推測することであり、その特性値は例えば商品の売上げ高や売り上げの個数などのように多くの場合定量的なものである。判別は、各個体を既知で排他的な複数個のカテゴリーのいずれかに分類することで、目的とする特性値が定性的なカテゴリーとなったのみでその目的は予測に類似である。多変量統計解析においては、予測では線形重回帰やロジスティック回帰あるいはポアソン回帰のように回帰分析としてくくられるものであり、判別では線形判別分析やロジスティック判別などとして方法論が提供されている。これらの統計手法は、入力を $x_1,...,x_p$ とし、出力を y としたとき、

$$y = f(x_1,...,x_p) + \varepsilon \tag{1}$$

とモデル化される。ここで ε は、$f(x_1,...,x_p)$では表現しきれない偶然変動項と見なされるもの（正確に言うと分析者がそう見なすもの）で、線形重回帰分析などでは正規分布に従

うと仮定されることが多い。いずれの手法も、入力変数の情報処理により最適な出力を得るための方法論と認識されよう。

［ⅱ］の予測・判別と［ⅲ］の因果関係とを混同してはならない。予測や判別では、与えられた説明変数 $x_1,...,x_p$ に対する y の値を推測するのみであり、説明変数の値を人為的に変えたときに何が起こるかまでは保証の限りではない。たとえば、値段設定(x_1)やスペック(x_2)の異なるいくつかの商品があった場合、それらの異なるいくつかの商品を元に、説明変数x_1、x_2から商品の売上高(y)を予測するモデルとして

$$y = b_0 + b_1 x_1 + b_2 x_2 \tag{2}$$

を作成したとき、説明変数の値が(x_1,x_2)である商品の売上げ予測は（2）によってできるであろうが、ある特定の商品の値段設定を変えた場合にその商品の売上げがどうなるかは分からない。あくまでもモデル（2）は異なる商品のデータから作成されたものであり、同じ商品の変数の値を変化させたときの結果までは保証していないのである。

また、(2）の係数b_1、b_2の値の解釈が可能かどうかという問題もある。たとえばb_1の値は大きくてb_2の値は小さいとき、値段設定x_1の y への寄与のほうがスペックx_2の寄与よりも大きいと言い切ってよいかどうかは議論の分かれるところである。予測の観点からは、予測の良しあしがある意味すべてであり、b_1、b_2がいかなる値を取るかはあまり関係がなく、解釈がなされないことも多い。その意味で（2）の関数はブラックボックス的であるともいえよう。

現在提供されているAIあるいは機械学習の方法論やシステムの多くは、この予測・判別を目的としたものである。AIブームのきっかけとなった囲碁・将棋のAIシステムは、現在の局面を入力とし、その下での最善な出力（着手）を与えるという意味でこの範疇に入る。モデルは（1）であり、機械学習の言葉で言えば y を教師とした教師付き学習と呼ばれるものである。AIあるいは機械学習の特徴は、(1）における変量の個数 p が膨大であること、および関数 f が非線形かつフレクシブルなことである。またディープラーニングをはじめとする多くの場合､関数 f は陽には表現されずブラックボックスとなっていて、ε が明示的には認識されない点も特徴といっていいであろう。予測あるいは判別のシステムのパフォーマンスはその的中度で測られるが、近年の多くの研究によると、AIおよび機械学習のパフォーマンスは、既存の多変量解析の各手法をしのぐものとなっているようである。

因果関係の確立［ⅲ］は、多くの科学的研究あるいは業務での究極的な目的であると言っても過言ではない。これを［ⅱ］の予測・判別と峻別することが重要である。因果関係が確立されれば、(1）における $x_1,...,x_p$ を人為的に変えることにより出力の y を変えることができる。すなわち（2）において、スペックx_2を変えずに値段設定x_1を1単位変化させれば、売上 y は平均的にb_1だけ変化すると言える。

因果関係の確立のためのゴールドスタンダードは実験研究であると言われる。新薬開発の臨床試験や次節で取り上げるWEBマーケティングのA/Bテストなどは実験研究である。

しかし、実験が倫理面やコスト面での理由で実行できない課題は数多く、その際は観察研究に頼らざるを得ない。観察研究からの因果推論は、近年の理論的進展と応用分野の広がりが顕著であり、第4節および最後の文献解題の節で改めて取り上げる。

次節ではA/Bテストを例にとり、それがなぜ有用であるかあるいは必要とされるかを統計的因果関係の観点から解きほぐす。

3. A／Bテストはなぜ有用か

A/Bテストとは、インターネットマーケティングにおいて、WEBページの何らかのよさ、たとえばページビューやクリック率の増加や究極的にはそれに伴う商品の売上げの増加を、AおよびBという2つのパターンのページを用意し、それらの比較に基づきより有効なWEBページを選択しようとするものである。その有用性が認識され、実際に実施されて効果を上げている例は多い。ここではこのA/Bテストを例に、統計的因果推論の鍵となるコンセプトと共にその有用性の源泉を詳しく見ていく。

3.1 因果推論における用語と記号

まず、統計的因果推論における用語を整理し、記号を定義する。因果関係の原因系をここでは処置（treatment）と総称し、結果系を結果変数（outcome variable）と呼ぶ。そして立証すべき対象を効果（effect）という。すなわち、ある処置が結果変数に与える効果を立証しようとするものである。処置もしくは結果変数は連続量あるいは個数のような量的な場合とカテゴリーで表される質的な場合とがある。A/Bテストでは、処置はWEBページの2つのパターンというカテゴリカルなものであり、結果変数はページビューの回数で効果は増加率という量的なものとなる。新薬開発の臨床試験では、処置は、試験薬の投与の有無であればカテゴリカルであり、薬剤の投与量であれば連続的となる。結果変数は、疾病が治癒したか否かであればカテゴリカル、血圧の効果量などであれば量的である。

ここでは簡単のため、処置は2種類の値を取るカテゴリカルなものであるとし、処置の種類を表す変数をZとし、Z=0もしくは1とする。新規の処置もしくは処置ありをZ=1、標準的な処置もしくは処置なしをZ=0としておくのが好都合である。たとえばA/Bテストでは新規開発ページをZ=1、既存ページをZ=0とし、臨床試験では新薬投与をZ=1、既存薬投与をZ=0とする。結果変数はYで表す。また、ZとY以外に観測される第三の変数をXとする。Xは、Zに影響を与えるがYに影響を与えない、あるいはZに影響を与えないがYに影響を与える可能性があるとき共変量（covariate）といい、ZとYの両方に影響を与えるとき交絡変数（confounding variable）という。さらに、XがZから影響を受けるときは中間変数（intermediate variable）という。ここでは第三の変数Xの呼称を変えているが、その役割を明確に認識することが因果推論の議論では肝要である。

3.2 比較の重要性

比較無しには処置の評価は望むべくもない、とは統計学の教えるところであり、真の因果関係の確立のために心せねばならない点である。新しいWEBページを作成したところページビューが増えた。ゆえに新しいWEBページは効果があった。薬を服用したら病気が治った。だからその薬には効果があった。ある教材を使って勉強したところテストの点数が上がった。よってその教材には効果があった。これらは日常的によく耳にする物言いであり、確かに効果があったかもしれないが、ないかもしれない。要は分からないということである。比較対照がないためである。新薬開発の臨床試験では、必ず新薬の比較相手として既存薬あるいはプラセボを置く。A/Bテストは、この比較の重要性をAとBという2つのパターンの設定によって実現している。その意味で、因果関係確立の第一歩はクリアしていることになる。

統計的因果推論における比較について述べておく。ある個体 i において、処置を受けたときの結果変数の値を $Y_i(1)$ とし、処置を受けなかったときの結果変数の値を $Y_i(0)$ とする。そして、個体 i の処置効果を

$$\tau_i = Y_i(1) - Y_i(0) \tag{3}$$

と定義する。ここでは $Y_i(1)$ と $Y_i(0)$ の差で効果を定義したが、比 $\omega_i = Y_i(1)/Y_i(0)$ で定義される場合もあろう。その場合は対数を取って $\tau_i = \log\omega_i = \log Y_i(1) - \log Y_i(0)$ とすればよい。個体 i は処置に対する反応に関してこれらの値の組 $\{Y_i(1),\ Y_i(0)\}$ によって特徴付けられる。この値の組を個体 i の潜在的結果変数（potential outcomes）という。$Y_i(1) = Y_i(0)$、すなわち処置の有無にかかわらず結果が同じであれば、その個体 i にとってはその処置は効果を持たないことになる。ところが、$Y_i(1)$ と $Y_i(0)$ の両方を観測することはできない。すなわち、個体 i は潜在的な値として $\{Y_i(1),\ Y_i(0)\}$ を持つが、処置の有無によっていずれか片方しか観測されず、もう片方は観測されない。それゆえ（3）は定義できても観測できない量となる。このことは統計的因果推論の根本問題（fundamental problem in causal inference）と呼ばれる。

3.3 処置効果の定義

3.2項で各個体の処置効果は推定不能であると述べた。では、推定すべき処置効果は何であろうか。それは母集団全体での処置効果の平均（平均処置効果 Average Treatment Effect = ATEもしくは平均因果効果 Average Causal Effect = ACE）であり、

$$\tau = E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \tag{4}$$

で定義される量である。ここで、$E[Y_i(1)]$は母集団の全員が処置を受けたときの平均、$E[Y_i(0)]$ は母集団の全員が処置を受けなかったときの平均であり、このままでは推定不能である。それを可能にする手立てが次の3.4項で導入するランダム割り付けである。

処置効果としては母集団全体での効果（4）以外にも、母集団のある部分集団を G としたとき、

$$\tau_G = E[\tau_i \mid G] = E[Y_i(1) - Y_i(0) \mid G] = E[Y_i(1) \mid G] - E[Y_1(0) \mid G] \quad (5)$$

とし、これを部分集団Gにおける平均処置効果とすることもある。部分集団としてはたとえば性別や年齢層などあらかじめ明確に定義できるものもあれば、処置を受けた人の全体とする場合もある。その場合は処置を受けた人での平均処置効果（Average Treatment Effect on Treated = ATT）ともいわれる。

ここで、母集団とは何かを明確にしておかなければならない。教科書的には、母集団はたとえばある大学における学生全体とかある市における有権者全体のようにあらかじめ定められていて、そこからのランダムサンプリングによってその母集団全体での平均（母平均）などを推測するという枠組みで話が進められている。しかし現実にはそうでないことも多い。母集団から得られたデータを分析するのではなく、集められたデータを一般化する対象として母集団を定義することも考えられる。

A/Bテストでは、明らかに当該WEBサイトにアクセスする人だけがデータ収集の対象となる。したがって母集団は「そのサイトにアクセスする（可能性を持った）人全体」とすべきである。もちろんそれはもっと大きな母集団全体の一部であって偏りを持つかもしれないが、ある商品を販売するサイトであれば、その商品に関心がなくサイトにアクセスしないような人は考慮の対象から外れても何ら差し支え無い。そのように母集団を定義する、あるいは明確に意識することにより、データ分析の結果をことさらに広い対象にまで一般化する愚を犯すことは避けられる。

3.4 ランダムネスの効用

比較が行われればそれで十分という訳ではない。比較対象がどのように選ばれ、処置が誰に施されたのかの情報が根本的に重要な要素となる。実験研究では実験参加者（被験者）の選択および彼らへの処置の割付けが研究者の手で行えることから、たとえば新薬開発の臨床試験では、臨床試験の適格者に対し新薬を投与するか既存薬を投与するかをランダムに決めている。処置のランダム割付けにより、サンプルサイズがある程度大きければ、処置を受ける群（処置群）とその比較対照の群（対照群）とで、処置の有無以外の条件がほぼ同じになることが保証される。したがって、処置群と対照群とで結果変数の値が異なったとすれば、それは処置の効果であると明確に判断することができる。それに対し観察研究では、自分の受ける処置の種類を被験者自らが決めているため、仮に処置群と対照群とで差が見られたとしても、それが処置の効果によるものかあるいは被験者の自己選択の結果によるものであるかの判断が難しい。

A/Bテストでは、両パターンへのアクセスを恣意的に選択しているのでなければ（これは実際問題として困難であろう）、パターンの選択はランダムと見なされることから、この条件をクリアしている。したがって、両パターンで見られたページビューなどの差異は、パターンの違いという処置効果によるものであるとの判断が可能となる。

統計的データ解析におけるランダムネスの効用としては、恣意的な観測対象の選択によ

る偏りの排除や、ランダムネスに基づく確率計算の源となることが挙げられる。それに加え、統計的因果推論では、以下で述べるようにランダムネスが本質的に重要な役割を担っている。たとえば処置の割付けがランダムであれば、その割付けに影響を与える変量は存在せず、3.1項で述べた第三の変数Xは常に交絡変数ではなく共変量として扱うことが可能となる。また、処置群と対照群とで観測可能な背景因子の分布を同じにすることは恣意的な処置の割付けであっても可能であるが、観測されない背景因子の分布までを同じにするためには、ランダム割付けが唯一の選択肢である。

第 i 個体への処置の割付けをZ_i=1（処置）=0（対照）としたとき、第 i 個体の観測される結果変数は

$$Y_i = Z_i Y_i(1)+(1-Z_i) Y_i(0) \tag{6}$$

と表される。このとき、処置群と対照群との観測される平均の差は、処置群の個体数をmとし、対照群の個体数をnとしたとき、(6) の観測される結果変数により

$$T=\frac{1}{m}\sum_i Z_i Y_i-\frac{1}{n}\sum_i (1-Z_i) Y_i \tag{7}$$

と求められる。処置の割付けがランダムであるときは$E[T]=\tau$が示される。すなわち、3.3項で「このままでは推定不能」とした推測対象の処置効果τが、割付けがランダムであれば観測された結果変数の平均を用いた(7)のTによって偏りなく推定されるのである。

以上見てきたように、A/Bテストは、統計的因果推論の立場から見たとき、処置効果が偏りなく推定できるという意味できわめて望ましい分析法である。母集団全体での処置効果だけでなく、(5) で示した部分集団での処置効果の推定も有益な情報を与えてくれることがあろう。有用な分析手法は、ここで見たように理論的な裏付けもしっかりしているのである。

4. 因果推論のテクノロジー

第3節で論じたA/Bテストは第2節で示した分類の中の実験研究の範疇に入るものである。実験研究であれば因果関係の確立がある程度容易になされる。ある程度としたのは、注意深く計画された実験研究であっても、データの欠測や割付けられた処置を遵守しないノンコンプライアンスなどが生じるためであり、その際には別途相応の対処が必要となる。また、第2節で述べたように実験研究は常に可能とは限らず、処置効果の立証を観察研究に頼らざるを得ない課題は数多い。

観察研究に基づく因果推論では、処置の割付けと結果変数の両方に影響を及ぼす交絡変数の除去が難しく、処置群と対照群の間でシステマティックな差異が生まれる可能性が否定できない。その場合は、処置群と対照群間で結果変数の値が異なったとしても、それが処置の効果であるとは断言できないことになる。そこで、観察研究であっても、交絡因子

の影響を排除して両群間でのシステマティックな差異を解消する手立てが必要となる。そのために開発されたテクニックが傾向スコア（propensity score）による調整である。

傾向スコア$e(X)$は、処置の割付けを表わすダミー変数をZとし、観測された共変量の集合をXとしたとき、Xが与えられた下で個体が処置に割付けられる確率

$$e(X)=P(Z=1\,|\,X) \tag{8}$$

として定義される。共変量Xは通常多次元であるが、傾向スコア$e(X)$は（8）の定義で分かるように確率という1次元の値である点が重要である。傾向スコアに関して次の2つの条件が想定される。

条件1（共変量の条件付きでの独立性）：観測された共変量X以外に割付けZに影響を与える変数はなく、潜在的な結果$\{Y(1), Y(0)\}$と割付けZは条件付き独立となる。すなわち$\{Y(1), Y(0)\}\perp Z\,|\,X$が成り立つ。

条件2（条件付き正値性）：与えられた共変量Xの下で、処置に割付けられる確率は0または1にはならない、すなわち$0<e(X)<1$である。

上記の2条件の下で、傾向スコアについて次の2つの性質が成り立つ。

性質1（傾向スコアの条件付きでの独立性）：傾向スコア$e(X)$が与えられたとき、潜在的な結果$\{Y(1), Y(0)\}$と割付け変数Zは条件付き独立となる。すなわち

$$\{Y(1), Y(0)\}\perp Z\,|\,e(X) \tag{9}$$

が成り立つ。

性質2（バランシング）：同じ傾向スコア$e(X)$の値に対応した個体の共変量の分布は処置群と対照群で等しい。すなわち$e(X)$の条件付きでXとZは条件付き独立

$$X\perp Z\,|\,e(X) \tag{10}$$

である。

これらの条件と性質に若干の補足を加える。条件1は、処置の割付けに関する情報はすべて観測される共変量Xが担っているという条件である。したがって、共変量Xが同じ個体同士では処置を施す前の背景因子に関しては両群間で差がないことになる。条件2は両群間で比較が可能となる相手が存在するという条件である。$e(X)$=1あるいは0となるXでは、片方の群にその比較相手がいないことになるので、比較可能性の観点から設定される条件である。

性質1および性質2は重要な結果で、傾向スコアの有用性を示している。実験研究では処置のランダム割付けが可能であり、この条件は潜在的な結果と割付けとの独立性$\{Y(1), Y(0)\}\perp Z$として表現される。観察研究ではその保証はないが、傾向スコアの値が同じ個体に限ってみれば、割付けがランダム、すなわち（9）が成り立つことが保証される。また、

処置効果の評価のためには処置を施す前では両群間での背景因子に差がないことが必要不可欠であるが、共変量Xは通常多次元であるので、個々の共変量の分布を同じにそろえることはきわめて困難である。ところが性質2の（10）は、傾向スコアという1次元の量が同じ個体では､すべての観測される共変量の分布が両群間で同じになることを主張している。

性質1および2は傾向スコアの有用性を示しているが、それは先に述べた2条件の成立下の話である。性質2はともかく、性質1の成立如何は立証が困難である。観測されていない交絡要因が存在する可能性は常に否定できない。そのため、考え得る共変量はなるべく多く傾向スコアの計算に用いたほうが良いという主張がなされている。観測の有無によらず共変量の分布を両群間で同じにするためには実験研究としてのランダム割り付けが必要となる。観察研究の限界と言えよう。傾向スコアの推定には、それが確率の推定であるためロジスティック回帰が用いられることが多いが、近年ではさらにフレクシブルな機械学習の手法の適用が提案されてもいる。

傾向スコアの利用法としては、マッチング、層別、逆確率重み付け法、および共分散分析が提案されている。これらについて簡単に述べておく。マッチングは、与えられた処置群および対照群の中から傾向スコアの値が同じあるいは類似の個体を選び出す手法である。すなわち全体の中から比較可能となる部分集団を作ることを目的とする。それに対し次の3つは与えられたデータをすべて用いる手法である。層別は、処置群および対照群を傾向スコアの値が類似であるような層に分け、各層同士では比較可能なようにする手法である。逆確率重み付け法は、両群の個体を傾向スコアの逆数で重み付けし、全体として両群間での比較可能性を担保する方法で、標本調査法による逆確率法を応用している。ダブルロバストネスという性質を付与する提案もあり、近年ホットな研究が続いている手法である。最後の共分散分析法は、傾向スコアを調整すべき重要な共変量に加えて共分散分析を行う方法である。いずれの方法も傾向スコアの1次元性を利用したものとなっている。

傾向スコアの利用以外にも近年、実験研究における処置の遵守違反（ノンコンプライアンス）、欠測データの処理、中間変数がある場合の因果パスの推定など、因果関係の確立に伴う統計分析で有用な手法が提案されている。これら統計的因果推論の理論の整備に加え、実際問題への適用を企図したソフトウェアの開発も進み、分析のためのテクノロジーは面目を一新しつつある。今後これらの手法を実際問題に適用し、さらにそれをブラッシュアップする必要がある。

5. おわりに

データサイエンスは、ビッグデータからの価値創造と称されることもあり、集めるというより集まっているあるいは集まってくる多種多様で大容量のデータがその分析対象であるとの認識がある。一方、統計的因果推論は、処置のランダム割付けといったデータの集め方が本質であり、比較的少量の質の良いデータをその分析対象としていて、第1節で述

べたようにビッグデータ解析の対極にある方法論と見なすこともできる。

第2節で述べたように、データ分析の目的は、現状把握に始まり予測や人為的介入の効果を知ることであることが多く、その際には因果関係の確立に関する確かな知識が必要となる。統計的因果推論の方法論を適用しなくてはいけない、というのではなく、それをひとつの規範あるいは理想像とし、そこからの乖離の程度を測りながらビッグデータ分析に挑むという姿勢が重要である。データが多ければすべてが解決するというのは幻想でしかなく、やはりそれらの集め方や集まり方に関する情報を収集し、それを活かすことが肝要である。

【文献解題】

統計的因果推論に関する知識を深めたい諸氏のため、単行本を中心に基本的な文献を紹介する。日本語で書かれた文献は残念ながらそう多くはない。本稿執筆では岩崎（2005）を参考にしたが、星野（2009）、宮川（2004）もそれぞれの視点から因果推論を論じた良書である。大森他（2013）、木原・木原（2013）は、分野が経済学および医療ではあるが、それぞれが定評あるテキストの翻訳であり、記述は分かりやすく得るところも多い。黒木（2009）も世界的に評価されている書籍の翻訳であるが、英文の原本も日本語の翻訳もやや難解である。

海外に目を転じると良書の出版が続いている。第一には、因果推論の立役者でそのモデルがRubin Causal Model（RCM）とも称されるD. B. Rubinと共同研究者のG. W. Imbensによる Imbens and Rubin（2015）が挙げられる。本書のタイトルにはAn Introductionとの副題があるが、本書の後に続く話題は数多いことからIntroductionと言えなくもないが「入門本」では決してなく、心して読む必要がある。また、Rubin（2006）はD. B. Rubinの因果推論関係の論文集であり、Rubin自身による解説も収録され、オリジナルの論文に当たるには便利な書物である。黒木（2009）の原本であるPearl（2000）は第2版Pearl（2009）が出版されている。Angrist and Pischke（2009）とKatz（2010）は共に翻訳が出版されているが原著そのものも読みやすく有用である。傾向スコアの生みの親の一人であり、観察研究の第一人者であるP. R. RosenbaumによるRosenbaum（2002）、Rosenbaum（2017）も座右に置くべき書物である。

その他、Gelman and Hill（2007）、Guo and Frazer（2015）、Holms（2014）、Morgan and Winship（2015）、Shadish, Cook and Canmbell（2002）などもそれぞれ持ち味を生かした特徴的で定評のある書物である。ハンドブックであるMorgan（2013）は情報源として有用である。Morton and Williams（2010）は政治学・政策科学における実験的手法を扱った書物で、米国などでは、この分野にも実験的な手法が適用されることを知ることができるという意味で興味深い書物である。

ここで挙げた以外にも因果推論関係の書物は多く出版されている。また、教科書的な書物をはじめとした最近の統計学の書物では因果推論的な扱いの占める割合が大きくなりつつある。ここで挙げた書物はどれも、眼下の流行に左右されることなく、大学の学部上級あるいは大学院において、じっくり勉強しながら読む価値は十二分にある。

【謝辞】

本学会へご紹介いただいた慶應義塾大学理工学部の鈴木秀男教授、および本稿の執筆の機会を与えていただいた学会誌編集委員の方々にお礼申し上げる。

【参考文献】

岩崎　学（2015）『統計的因果推論』朝倉書店。

大森義明・小原美紀・田中隆一・野口晴子（訳）（2013）『「ほとんど無害」な計量経済学－応用経済学のための実証分析ガイド－』NTT出版（Angrist and Pischke（2009）の邦訳）。

木原雅子・木原正博（訳）（2013）『医学的介入の研究デザインと統計－ランダム化／非ランダム化研究から傾向スコア、操作変数法まで－』メディカル・サイエンス・インターナショナル（Katz（2010）の邦訳）。

黒木　学（訳）（2009）『統計的因果推論－モデル・推論・推測－』共立出版（Pearl（2000）の邦訳）。

星野崇宏（2009）『調査観察データの統計科学－因果推論・選択バイアス・データ融合－』岩波書店。

宮川雅巳（2004）『統計的因果推論－回帰分析の新しい枠組み－』朝倉書店。

Angrist, J. D. and Pischke, J.-S.（2009）*Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Gelman, A. and Hill, J.（2007）*Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press.

Guo, S. and Fraser, M. W.（2015）*Propensity Score Analysis: Statistical Methods and Applications, Second Edition*. Sage Publications.

Holms, W. M.（2014）*Using Propensity Scores in Quasi-Experimental Designs*. Sage Publications.

Imbens, G. W. and Rubin, D. B.（2015）*Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Katz, M. H.（2010）*Evaluating Clinical and Public Health Interventions: A Practical Guide to Study Design and Statistics*. Cambridge University Press.

Morgan, S. L.（Ed.）（2013）*Handbook of Causal Analysis for Social Research*. Springer.

Morgan, S. L. and Winship, C.（2015）*Counterfactuals and Causal Inference: Methods and Principles for Social Research, Second Edition* Cambridge University Press.

Morton, R. B. and Williams, K. C.（2010）*Experimental Political Science and the Study of Causality*. Cambridge University Press.

Pearl, J.（2000）*Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J.（2009）*Causality: Models, Reasoning, and Inference, Second Edition*. Cambridge University Press.

Rosenbaum, P. R.（2002）*Observational Studies, Second Edition*. Springer.

Rosenbaum, P. R.（2017）*Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press.

Rubin, D. B.（2006）*Matched Sampling for Causal Effects*. Cambridge University Press.

Shadish, W. R., Cook, T. D. and Campbell, D. T.（2002）*Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company.