ビッグデータ時代の統計学教育の在り方について

天野　徹
Email: amano@soci.meisei-u.ac.jp

明星大学人文学部人間社会学科

◎Key Words 統計学の教養教育，仮想母集団，リスク管理

1.　問題関心

文系の学生に対する統計学教育は、どうあるべきか。学生時代、統計学の授業を受けていた当時から、この疑問は長い間、私の頭の中にあった。大学院で標準化調査を重ね、データの分析・管理にかかわりながらも、この問いはずっと、私の頭から離れなかった。大学に職を得、統計学教育にかかわるようになってから、学生たちを教えるのに相応しいと思われる教科書を探したが、納得のいくものは見つけられなかった。社会学系の学生向けの統計学の教科書として、ボーンシュテット＆ノーキの「社会統計学」の訳本がハーベスト社から出て数年たっていたが、数学が苦手な学生たちに教えるには、まだなお難しすぎたからである。

統計学とか数学というものにも色々流儀があって、系統が違えば記号の体系が違ったりすることが珍しくない。確率分布にしても検定量の分布にしても、どうしてそういう式で表されるようになったのか、そもそもどうして近似しなければならないのか、その理由がわからない。「自分の頭で納得できなければ、使うことができない」ということであれば、表記法から概念に至るまで理解不能で、使い方についての根拠もわからないものを、どうしてノウハウとして身に着けられようか。学生たちの反応をそう解釈した時から、「自分が学生時代に受けたかった講義内容を、自分で考えて教科書を作ろう」と考えて以来、文系の学生である私の、苦難に満ちた研究が始まったのだった。

当時、桃山学院大学におられた安藤洋美先生をはじめとして、様々な方々にお世話になりながら、この試みは書籍の形にまとめることができた。２項分布をコイン投げから説明し、順列・組み合わせ、２項定理、ド・モアブル＝ラプラスの定理を経て、正規分布による近似を説明し、誤差および外れ値の管理から誤差分布探求の歴史をたどり、最小二乗法を巡る争いから正規分布の第三証明を経て、その式の成立の理由や、記号法について説明する。正規分布に関連する概念の誕生の経緯と、２項分布や正規分布による社会認識の歴史について触れた上で、標本抽出と基本的検定手法および、その論理的な基礎となっている統計的帰謬法について説明し、第三変数を活用したエラボレーションの方法について解説する。最後に、探索的データ解析の手法について説明し、正規分布に近似できないデータについても、基本的な分析ができるよう、配慮した。

現在も私は、この本を、社会統計学や標準化調査の教科書として用いており、学生からの評判は悪くない。講義の後で学生から積極的な質問が寄せられることも珍しくないし、拙著を使いながらの解説で文系の学生も理解にいたり、喜んでくれる。期末試験の直前に「補講」を希望した学生を相手に、午後６時から１時間半をかけて行った特別授業では、学生たちから出される質問すべてについて、確率統計の思考方法から理論・公式の成り立ち、計算方法などについて解説を行ったが、講義終了時には学生たちが「もう７時半なのか。とても楽しかった。こんなに時間がたつのが速いなんて!!」という言葉が発せられたのには、驚いた。数学が苦手と思い、確率統計など理解できるはずがないと思いこんでした文系の学生たちも、心の底では、自分の納得できるように理解した上で、そのスキルを使えるようになりたいと望んでいたのである。そう、彼らは「統計学へレンケラー」だったのだ。

さて、ビッグデータの時代を迎え、データサイエンティストの育成の必要から、日本の統計学教育も、大きく変わろうとしている。統計学が学問として成立するのかが議論された時代を知っている身としては、「統計学が最強の学問である」というタイトルがついた本が出版されたということに、一種の感慨さえ感じる。また、統計学部のない国の悲しさか、統計学についての歴史的な考察を、学生たちの統計学教育のためにまとめなおす方が、ほとんどいないということに、逆に違和感を感じたりするから不思議である。ビッグデータの時代において、統計学教育が、数学的なモデル構築と検証による意思決定の質の向上を目指すのは、当然なことだ。しかし、統計学教育の改革の内容を見る限り、そのモデルの数学的なイメージについてきちんと理解させようとしているか、そして意思決定時の「もうひとつの」重要事項である、母集団の状態の推定とリスクの管理について、きちんと理解させようとしているかは、疑問である。なぜなら、従来の統計学の教科書について感じた「推測統計における検定量の算定の論理についての教育」が、必要十分な程度に改善されているとは思われないからだ。

情報ネットワークと計算機の高度化により、統計的データ処理にビックデータとビジュアライゼーションが加わったことで、データ活用の可能性は大きく広がった。モデルの構築とあてはまりの良さに関して、直感的に判断できるような環境ができたからである。それと同時に、意思決定を誤る危険(リスク)を管理する必要性も大きくなっているが、「統計学の参照基準」を見る限り、検定のロジックや検定量の意味についての教育は、十分になされているとは思えない。ビックデータを分析する際にも、意思決定のリスクを管理するた

めには、仮想母集団の想定が必要であり、この概念を使いこなすには、検定量を算出するロジックについての理解が必要である。また、統計による意思決定のリスクを理解するためには、統計学の社会史についての教育も必要なのではないか。特に文系・社会科学系の領域においては、対象とするデータが社会に関するものであるから、それに統計的手法を適用することの意味を理解する必要があると思う。統計学教育の改革は、これまでの「数学が苦手な文系学生を切り捨て、頂点だけを高くする」のではなく、「すそ野を広げて、これまでにない多様な展開を、新たな可能性を志向するもの」であって欲しいと思う。

本稿では、このような立場から、ビッグデータ時代における推測統計の意義について再考し、特に社会学を専攻する学生を想定して、大学における「統計学教育の参照基準」では触れられていないものの、これからの大学での統計学教育で必要と思われるポイントを指摘した上で、その解決法を提案することとしたい。

2. 文系学生に対する統計学教育の現場と参照基準の検討

まず、平成 26 年 8 月 1 日に公表された「統計学の各分野における教育亭編成上の参照基準」から、社会学を学ぶ学生について関係する部分は、「大学基盤科目としての統計教育の参照基準」と「社会学分野における統計教育の参照基準」であると考えられる。ここではまず、それぞれの参照基準のポイントを抜粋して示し、検討してみることにしたい。

2.1 大学基礎科目としての統計教育の参照基準について

この参照基準では、まず最初に、統計学を「自然科学、人文科学、社会科学、生命科学のあらゆる学問領域において、データに基づく実証研究を科学的に行うための学問体系である」と位置づけ、「仮説の発見・構築や検証のための実験、調査、観察研究の過程で得られるデータに基づいて正しく推論を行う力は、すべての学問分野で必要とされている」と、その必要性について断じている。その上で、これからの大学が育成すべき人材について、「実験や調査・観察研究という研究の違いを認識した上で、適切なデータ収集法の理解と実践、得られたデータを要約し、グラフなどを用いて分かりやすく表現するスキルが求められる」と述べた上で、そのような人材になるためには、「母集団と標本、標本誤差の知識や不確実な事象の起こりやすさを表す確率や確率分布の知識の習得も求められる」という。

参照基準では、そうした人材を育成するために、、大学が彼らに身に着けてもらうべき能力として、「統計学の役割と公的データの活用能力」、「記述的統計解析スキル」、「推測的統計解析スキル」、「統計解析の結果判断能力と分析スキル」を挙げ、その具体的な項目として、①統計学の役割と活用事例、②データの要約とグラフ化(記述統計学的手法)、③研究の種類とデータ収集法、④確率と確率分布、⑤統計的推測、⑥コンピュータの利用を挙げている。こうした内容は、実はビッグデータ活用の時代以前から、統計学教育について当たり前に論じられてきたものであって、目新しいものは特にない。要は、コンピュータによるデータ処理能力の飛躍的な向上と、センサー技術等によるデータ量の爆発的な増加により、勘や経験に基づく意思決定を脱し、豊富なデータについて分析・視覚化した上での合理的な意思決定が可能となったことに、教育内容を適合させたい。難解な記号や数式、哲学や歴史、検定のロジックやリスク管理の方法をスキップしこれを大量データと高速コンピューティングによって代替させようということであろう。しかしながらこれは、これまで日本の統計学教育が怠ってきた、「教養としての統計学」を切り捨て、「スキルとしての統計学」に特化するということだ。こうした方法論は、数学が苦手な学生に「実践的」な教育をしているように見えて、その実、コンピュータというブラックボックスに入れて出てきた数値を、盲信するタイプの人材を大量生産してしまうことになるのではないか、と危惧する。

2.2 社会学分野における統計教育の参照基準について

次に、社会学分野における参照基準についてみよう。社会学分野における参照基準ではまず、「基礎科目としての統計教育に加え、大標本調査における諸々の調査と過程、自動記録によって収集された市場データの扱い方、および官庁統計の活用法とその限界などを正しく理解することが、当分野で習得すべき中心的内容となる。」としている。そして、専門課程における統計教育として「調査や実験によるデータ収集方法、調査データの分析手法、市場調査や官庁統計の活用についてのスキルなどを中心とする授業が配置されること」が必要であるとする。

参照基準では、そうした人材を育成するために、大学が彼らに身に着けてもらうべき能力として、「統計学の役割を理解する能力」「調査データ収集に関する能力」「記述的データ解析スキル」「推測的データ解析スキル」「探索的データ解析スキル」「統計ソフトウェアを活用し出力結果を解釈する能力」を挙げ、その具体的な項目として、①統計学の役割と活用事例、②調査データ収集に関する知識、③データの可視化と要約(記述統計学的手法)、④確率と確率分布、⑤統計的推測、⑥基本的データ分析、⑦コンピュータの利用、を挙げている。しかしこれらの項目は、この参照基準ができる以前に作成された社会調査士過程のカリキュラムにおける、統計学に関連する科目に、既に含まれている。

〔社会調査士課程における統計学関連の科目内容〕
【Ｃ】 基本的な資料とデータの分析に関する科目
【Ｄ】 社会調査に必要な統計学に関する科目
【Ｅ】 量的データ解析の方法に関する科目
(注：社会調査協会ホームページ
http://jasr.or.jp/participation/curriculum_sr.html を参照)

このカリキュラムと参照基準とを比較してみれば、参照基準は社会調査協会の提示するカリキュラム内容を整理し直したものにすぎず、新たにビッグデータ分

析に対応したものでもなければ、これを機会に新しいスキルを身につけさせようという目的をもってつくられていたわけでもないことがわかる。目新しいのは「自動記録によって収集された市場データの扱い方」という文言だけで(おそらく、センサーデータの大量発生を意識してのことであろうが)、それを分析する環境やスキルが具体的に示されているわけではないからだ。

3. 参照基準を超えて―社会学を専攻する学生への統計教育の改善のために―

さて、それでは、大学基礎科目としての統計学教育は、どのように改善すべきだろうか。そして、社会学を専攻する学生を対象とした教育は、どのように改善すべきだろうか。

3.1 大学基礎科目としての統計教育について

スキルの教育はもちろん非常に重要である。しかしながら、大学教育では、少なくとも、そのスキルの基になっている統計学の社会史について、教える機会を設けるべきではなかろうか。統計学が社会現象に応用された当初、社会に関するデータの分析に基づいて行われた政策の大半は、当初の目標を達成することができなかった。そうした事実を教えることなしに、データ分析のスキルを教え成功例だけを教えることが、果たして統計学教育を教えるうえで適切な方法といえるだろうか。

また、具体的な教育内容としては、データの要約とグラフ化があげられているが、ソフトウェアを使って「箱ひげ図」を描くことを教える前に、生のデータを人の手で視覚化する「幹葉表示」を教えるべきではなかろうか。まず最初に、人の目と手を使ってデータを処理することを体験させることこそが、統計学的なデータの理解において、非常に大きな意味がある。参照基準に上げられた項目は、ICT の活用にこだわるあまり、普通の学生がデータを理解する上で必要な「目と手を動かす」という経験を排除してしまっているように思える。

さて、文系の学生を対象にして統計学を講じる際の困難はまず、確率変数という概念および、近似の概念を理解させることの難しさに始まるといってよい。これらは確率分布と同様、確率統計の世界観に関することであり、統計学によってリスクを管理しながら意思決定を行う際の基本となるべき事柄であるが、そうしたことにきちんとした配慮がなされているか、疑わしい。最終的には p の値で判断できればいい、というような安直な発想でスキル教育に終始してしまうと、統計学を使いこなす人材ではなく、データから使われる人材が大量生産される危険があるのではないかと思う。

ビッグデータは「全数データ」であるといわれている。しかし、ビッグデータを分析することで得たい情報は、たいていの場合、「まだ現実となっていない未来」に関する予測である。とするならば、ビッグデータの分析において、推測統計学は新たな意味合いを持ってくることになる。従って、ビックデータの活用において、推測統計によるリスク管理を行う意義と、そのロジックについては、きちんと教えておく必要があるはずなのだが、参照基準を見る限り、そうした記述は特にない。これでは、新しい参照基準によって教育され、ビックデータを活用する立場になる人は、どのような論理でリスクを管理しているのか理解しないままに、リスクを管理した気持ちになっても不思議ではないだろう。

私の経験では、文系向けの確率統計の教育において、リスク管理という概念を理解する上で必要な、条件付確率や事前確率・事後確率、片側検定における事前情報の価値などについて、きちんとした教育ができているか、非常に疑わしい。したがって、こうした基本的なことが理解できない学生たちに、ビジュアライゼーションや仮設検定を教えることは、実は非常に危ういことではないかとも思われる。不確実な情報しか得られない場合に、あるいは、データに偏りがある場合に、様々な情報を吟味してリスクを管理しながらよりよい意思決定ができるようになることこそ、確率統計を学ぶ者の目指すべきところであるが、参照基準を見る限り、そうした想定はなされていないようである。

さて、これらのポイントは、文系理系を問わず重要なものであるから、大学基礎科目としての教科で、すべての学生に対して一通り教えるべきであると思う。これらの基礎がなければ、統計理論を誤用しても気が付かないし、データからの仮設の発見において不利であるし、標本の偏りについて鈍感になる可能性があるし、事前に得られているデータの評価が出来ずリスクのコントロールに過不足が生じる可能性がある。それが、「本当に分かっている人たち」との間で、意思決定の質に決定的な違いを生むことが、明らかだからである。

3.2 社会学分野における統計教育について

ビッグデータの時代を迎え、社会学的センスと確率統計のセンスを併せ持った専門家を育成するのであれば、少なくとも以下の点については、参照基準に書き加えておくべきではないかと思う。

まず、参照基準で示されている官庁統計の利用であるが、官庁統計は現実に実社会の社会システムの運営に生かされているのだから、社会統計が人間の社会認識に対してどのような影響を与えてきたか、そしてそれが妥当なものであったか、その結果として社会がどのように変化したかについて教えるべきではないのか。また、統計学を活用することによって社会システムの運営がどのように改善されてきたか、それが所期の目的を達成できたのか、社会に対しどのような影響を与え、新たな問題を生み出したか等、社会学ならではの統計学教育の内容を提示すべきである。

次に、層化抽出・確率比例抽出を行った場合の、標本の統計学的な意味と、検定時におけるリスク管理の違いについて、きちんと教えるべきである。また、層

化抽出・確率比例抽出を行った場合の、母分散の推定および検定量の算出方法について、可能な限り丁寧に教えて、標準化調査における標本の代表性の評価と標本誤差、検定の注意事項などについて、分析者が正しく認識できるような教育内容を明示すべきである。これらのロジックは、ビッグデータに対する推測統計学の応用においても、需要な意味を持つからである。

推測的データ分析においては、それぞれの検定手法のモデルを視覚的に解説した上で、検定の確率論的な基礎について、分かりやすく教えるべきである。例えば、分散分析における級内平均平方は、「各級内における偏差平方和から推定した母分散の値」、級間平均平方は「各級の平均値と標本全体の平均値との差から推定した母分散の値」という解説を、計算方法とともに示しながら教えるべきである。それを通して、統計的検定の危うさを学ぶことが、リスクの管理を意識しながら行っていく上で、有効と思われるからである。(級間平均平方を級内平均平方で割った値が F 値であると教えも、ほとんどの学生は検定のロジックを理解できないまま p の値を盲目的に判断するだけだろう。)

さらにに、正規分布に従わないデータを適切に分析するために、探索的データ解析の技術をきちんと教えるべきである。分布の形状とデータの散らばり具合および代表値、そして、外れ値などについては、データ分析についての記述の最初に、分かりやすく書くよう指導すべきである。人文科学に限らず、現実の社会には正規分布に従う確率変数はむしろまれなのであるから、そうしたデータの処理に必要な技術をきちんと教えるべきであるし、離散変数の分析に必要な対数の概念なども一通り教えるべきなのではないか。こうした基本的な教養の不在が、分析結果の解釈に悪影響を与える可能性があると思われるからである。

4. ビッグデータ時代の、文系学生向け統計学教育への提言

データサイエンスの基礎としての統計学の技術のポイントは、数値データの処理の方法とビジュアライゼーションによるモデル構築の手法になる。この点については、参照基準は、妥当な内容といえる。しかしながら、これらのポイントを突いただけで、統計学教育が改善されるとは思われない。

その最大の理由は、文系の教員・学生は、理系の教員・学生と異なり、「失敗」ということに慣れていないことにある。文系での統計学教育では、文系と理系のカルチャーの違いについての注意が必要である。つまり、ビッグデータに対応した統計教育を機能させるためには、問題発見から問題解決の例題、問題未解決の例題などを、歴史的エピソードとともに教えて、データ分析はトライアル&エラーが当たり前であることを実感させる必要がある、ということだ。

また、理系が技術として統計学を学ぶことに慣れているのに比べ、文系の学生のほとんどは慣れていない。従って、同じように統計学の技術を教えても、文系と理系では受け取り方・使い方が大きく異なるものとなる。そして、この違いを乗り越えるためには、文系理系を問わず、統計学教育の最初に、統計学を思想や哲学、そして社会史の文脈で教えることが必要である。こうすることで初めて、文系の学生にとっても統計学の重要性が肌でわかるようになるだろうし、一つ一つの研究で使われる技術が、実社会の中で歴史とともに発展し社会システムの高度化を実現してきた技術と通じていると実感できるようになるだろう。

長い間文系の学生を対象にして確率統計を講じてきた経験から言えば、文系の学生は、確率分布を表現する複雑な式を見るだけで拒否反応を示す傾向がある。確率分布のパラメーターとして使われる μ や σ そして σ^2、母平均の記号として使われる M など、始めて目にする記号の読み方や由来について、詳しい説明を求める学生も散見されたが、「σ と Σ は違う文字なのに、どうして同じくシグマと読むのか」とか、「σ^2 と s^2 と V はともに分散を示す記号だが、どうして記号が違うのに同じ量を示すのに使われるのか」などの問いに丁寧に答えていかなければ、文系の学生たちが統計学について持つ違和感を払拭することはできない。また、正規分布がなぜあのような形そして式で表現されるのか、どうしてわざわざ「正規」と呼ばれるのか、その説明力はどの程度なのかなどについて、きちんと教えておかないと、誤用したことにも気が付かない似非専門家が跋扈することになる。

ビッグデータ時代は全数調査ができるから、推測や近似は必要ないと誤解されがちだが、ビッグデータの時代だからこそ、今まで以上に、正しい母集団の想定とリスクの管理が必要であり、そのためには、今までの統計学教育がないがしろにしてきた「統計学の教養」こそ、見直されるべきである。時代の要請に対応するための統計学のスキルは、いってみれば、水上に見える氷山の頂き部分のようなもの。その頂きは、水面下にあって見えない氷塊にあたる「教養」によって支えられる必要がある。教養のない表層的なスキルは、かえって小手先の技に終始し、操作されやすい人材を育ててしまうのではないか、とも思う。

統計学の参照基準に示された「氷山の頂き」部分を無にしないために、統計学の教育に携わる者は、水面下にあって注目されない「統計学の教養」授ける試み、そして努力を、行っていくべきである。そしてそうした教育を受けることによってこそ、学生たちは、ただ単にブラックボックスとして統計手法を濫用する「統計学のスキルに振り回される存在」になることなく、統計学のスキルを自分のものとして、自覚的に活用していくことができる「統計学のインテリジェンスを持ったヒューマンな存在」になれのではないか。そして学生たちの多くは、心の底では、そういう存在になりたいと願っているのだと、確信している。

参考文献

(1)イワン・ハッキング著、石原英樹・重田園江訳、「偶然を飼いならす―統計学と第二次科学革命―」、木鐸社、1999
(2)鎌倉稔成・田栗正章・西郷浩ほか、「統計学の各分野における教育課程編成上の参照基準」、統計関連学会連合理事会・統計教育推進委員会・統計教育大学間連携ネットワーク質保障委員会、2014
(3)天野徹著、社会統計学へのアプローチ、ミネルヴァ書房、2006