データ分析は科学か: 多変量解析編

○狩野 裕
大阪大学 大学院基礎工学研究科

1. 結論　「データ分析は科学か，それともアートか?」というお題をいただいた．実質科学 (substantial science) と方法論 (methodology) という言葉がある．実質科学とは解明・解決すべき具体的な現象・問題へのアプローチである．方法論は実質科学の目的を達するための方法を研究対象とする学問で，具体的な現象・問題が存在する場合もあるが，それを意識しないケースもある．一般論を言えば，これらは科学の発展における車の両輪であり不可欠なものである．統計学は実証科学における統計的方法論を提供する学問と位置づけることができる．それゆえ，科学の一翼を担っていることは疑いの無い事実である．

2. 議論　では，このような明白な事実があるにもかかわらずなぜ学会においてこのようなトピックの特別セッションが組織され議論することになるのか．筆者は過去にこのようなテーマが議論されたかどうか寡聞にして知らない．しかしながら，確かに，統計科学に携る多くの方々にとって興味を惹く話題のように思える．そして，このようなトピックを議論することに至った理由もいくつか思い浮かぶ．各論に入る前に科学の定義を見ておこう．広辞苑第五版によると「科学」の定義は次のようである．

> ① 体系的であり、経験的に実証可能な知識。物理学・化学・生物学などの自然科学が科学の典型であるとされるが、経済学・法学などの社会科学、心理学・言語学などの人間科学もある。② 狭義では自然科学と同義

2.1. 方法論について　まず最初に考えたいことは日本人の方法論に対する理解不足である．大学生の何パーセントが方法論という言葉を知っているだろうか．ここには二つの論点がある．一つは，明治開国以来，西洋文明に追いつけ追い越せとの号令の下，方法論を研究するよりは，その成果を得て実質科学のレベルを向上させることに力点がおかれたのではないかということである．この傾向は現在も続いており，時に「方法のタダ乗り」と批判されることがある．他の一つは，統計科学が境界学際領域的であることである．例えば，統計科学は応用数学であり数学と実質科学の境界と位置づけられることがある．いままで，日本では伝統的な学問が一流であって新領域や境界領域を重視してこなかったきらいがある．

統計学が科学でない（もしくは二流の科学）という議論があるとすれば，根底には方法論に対する，このような根も葉もない偏見があったのではないか．

2.2. 多変量解析における探索的方法について　多変量解析は複数個の現象を同時に扱う方法論であるから，一次元の統計的手法よりも適用が難しいことが多い．多変量解析の目的の一つに多くの現象の間に横たわる因果性，相関性，共起性の同定がある．統計的因果推論はそのための道具立てであり，近年，注目を浴びている研究分野の一つである．扱う現象が極めて多かったり（超高次元），動的なデータが逐次観測されたり (on-line sampling) するような状況から有益な情報を得ることもしくはそのための方法をデータマイニング（DM）という．

「マイニング」は必ずしも positive な用語ではない．広大な大地の中から金や銀，石炭の鉱脈を発見する事業はギャンブル的要素（山師？）があり，また，その発見は詐欺的（God Hand?）か超能力者によるもののようでもあったから，どうしても非科学的という響きが伴う．背後の理論に乏しくデータの情報のみに基づき計算機のパワーに任せて因果性，相関性，共起性を調べる方法には非科学的な匂いがするのも事実かもしれない．しかし，それはDMを適用する側の姿勢の問題であって，少なくとも，DMを実現する方法の開発やそのための理論整備は情報科学の重要な研究分野である．

2.3. The bumpy road to a model fitting　標題はある高名な心理学者のため息である．実験データであればデザインに合わせた分散分析等を行えばよく，勝負は早い．しかし，SEM 等で適合の良い因果モデルを探索し始めると体力勝負というわけである．また，探索的な分析には様々なノウハウが存在し，それらは経験によって，または，口伝によって獲得・伝承されている．統計学のテキストはノウハウの獲得に一定の役割を果たしているのは事実であるが，モデル探索などの探索的統計解析の「泥臭いプロセス」を十分に体系化して記述しているかは疑問

である．ときに，統計ユーザから，統計のテキストは「痒いところに手が届かない」という批判を受ける．この点については，自省の念に駆られる．データ解析術の体系化は，科学的なプロセスとしてのデータ解析をより客観的にする．

2.4. 潜在変数について

Factor analysis is an art!

これは，丘本正大阪大学名誉教授がある国際会議で因子分析についての研究成果を発表したときにフロアから得た質問である．同教授がこの質問にどのように対応したか興味があるが，残念ながら筆者は「何とか切り抜けた」という事実しか知らない．因子分析や潜在変数モデルなど直接観測できない概念を扱う多変量解析のモデルは多い．しかし，その妥当性について統計科学者でも批判的な立場をとる場合がある．多くの母数を伴う探索的因子分析モデルは，推定が難しいのは事実であるがそれは既に指摘されており，現在は，事前の情報をより多くモデルに組み込んだ検証的因子分析モデルが開発され，安定的な推定ができるようになっている．

統計モデルにおける潜在変数の導入には歴史がある．その代表は誤差変数であり，続いてブロック因子等の変量効果である．ただし，これらについては，潜在変数モデルにおける潜在変数のように，特定の概念を具体化するという意味づけはしていない．潜在変数（構成変数）は構成概念の数理モデルであり，潜在変数を規定する測定モデル (measurement model) は心理学の尺度構成の理論と方法に基づいている．それはモデルであるから完全ではないが，多くの統計学者は，尺度構成の理論を学べば，その矛先を収めるのではないかと思われる．

2.5. 物理学帝国主義について　物理学やそれを援用する工学等狭義の自然科学に代表される理論が直接適用できない状況において，とりあえずデータを採取し分析することで現象に対する知見を得るという作業に批判があるのではないか．以下はインターネットのあるページからの引用である．物理学をかじったことのある投稿者の統計的方法に対する偏見がダイレクトに表現されている．

> いちばん問題なのは＊＊[1]で問題を解決したから何なの?ということだ。ノウハウというのは物理的理由を理解してはじめてノウハウになるのであって＊＊[2]最適化はそのとき得られる結果の先には何もない。トロイ制御問題ばかりやってないで複雑な数理モデルで実験やってみなよ。最適値らしきものはみつかるかもしれないが、それは物理的に正しいか？ やってみればわかる。

この指摘が正しいかどうかは別にして，物理学や工学等を経験した人が，ややもすれば統計分析についてこのような見方をしている可能性がある．実際，統計分析だけが理論根拠となるような原理原則は再現性が乏しいかもしれない．それに物理的理由が付加できれば確固たる「科学」に昇華するかもしれない．問題は，現在において統計的な分析以外にどのような科学的アプローチが存在するのかということである．有効な物理的接近法が存在し適用できるのにそれを怠ることは科学的ではない．もしそうならば上述の批判は尤もである．統計的な接近に続いて物理的な方法によって確証すべきであろう．

当然ながら狭義の自然科学的方法は万能ではない．それらの方法を直接適用できない学問領域は多い．現在のところ，社会科学の多くの研究において狭義の自然科学的方法は適用できないのではないか．上述の投稿者は今まで物理的な接近法で解決できた問題ばかりを扱った経験しかないのではないか．社会科学に関連する実際の問題に何らかの解決策が求められたとき，上述の投稿者であっても統計的接近法を取ることがあるだろう．

だからと言って統計的方法を安易に適用するのはよくない．とりあえずデータを採るという姿勢は科学的ではない．research question に関する詳細な先行研究をレビューし，それらを活かして，母集団の規定，無作為抽出，無作為割当て等を含めたデータ採取のための計画と実施の方法を入念に検討しておかなければならない．統計科学者と実質科学者が協力すべきところであろう．

3. 終わりに　統計学は科学研究における方法論を提供すると同時に科学の文法でもある．この意味で統計学は科学の発展に一定の寄与をしてきたし，今後もし続けるであろう．しかしながら，中には非科学的であると批判される部分もある.「データ分析は科学だ！」と反駁するだけでなく，批判を真摯に受け止め，データ分析をより体系的な方法論として，また，不整合を最小化した科学の文法として完成させるための原動力にしたい．

発表当日では，これらの議論に加えて，実証科学の位置付け，evidence-based method 等にも言及しつつ，統計科学の方向性について議論したい．

[1] 多変量解析のある方法
[2] この分野のテクニカルターム