統計十話

第10話　探索的にデータを取扱うことの大切さ（No. 2）

—データの科学の方法論—

林　知己夫

文部省統計数理研究所

2.「データの科学」誕生の経緯

いわゆる数理統計学の行き方にあきたらず，「統計数理」を標榜して統計学の異端を目指し，データによる現象解析のあり方を研究してきた私どものグループは，その結果として，標本調査・多次元分析・数量化・分類などを中心とする成果を積み上げてきた．

これとは独立に「数理統計学」の行き方と真っ向から対決する姿勢を示して新しいデータ解析の方法を目指したフランスのベンゼクリを中心としたグループがある．両グループは日本学術振興会による日仏研究集会において相まみえたのである．共鳴するところが多く，その後，緊密な連係をとることになり，第一回日仏セミナーが1987年に東京で行われた．それ以後，いわゆる「データ解析」(data analysis, analyses des données) を発展させることが必要だし，このためには新しい概念を必要とするということになった．それを，Data Science と名付けることにした．

第二回日仏セミナーは Data Science をキャッチフレーズとして掲げ，1992年にモンペリエで開かれた．その概念化は十分ではなかったが，データに関する包括的な方法論を中心とすることにし，これまでにないものを作り出そうとする動きであった．

3．原点はデータによる現象解析

統計学にせよ，データ解析にせよ，出発点はすばらしいものであり，役に立ってきたことは周知のことである．しかし学問が分化し，進展してくると，研究が過度に数式的になり精密化してくると本来の目的が見失われてきて，沈滞し，活力が枯渇してくる．理由は，データを以って現象を解析する根本理念が忘却されることにある．現在のこうした状況を超えて，活力ある学問が発展してくるためには，統計学・データ解析・分類・その他の関連諸方法を統一する哲学が不可欠で，私はこれを「データの科学」と名付けたのである．

原点は「データによる現象解明」—こうした方法を作る側に立てば，主体的・能動的な表現でなければならないので「データによる現象理解」と言いたい—という点にある．全てこの一点に向けて，方法論・方法・理論・実施が行われねばならない．言い換えると，データの科学はデータを以って実際の現象を解析し理解することを志向し，統計学・データ解析・分類・その他の関連諸方法を統一する理念であり且つそれに基づいて生産される諸結果を包含するものである．

これまでの諸学問の成果を踏まえ，且つこれに囚われることなくポテンシャルとして活用し，複雑な自然・人間・社会現象の諸相，隠された構造を露呈させることが大きな目的

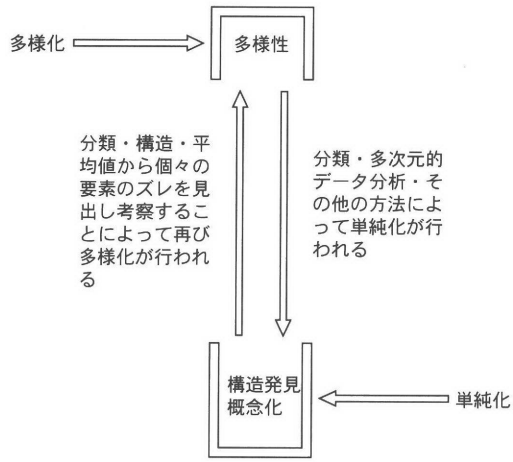
となる．比較的に単純な現象は伝統的方法で成果をあげることもできるが，従来の方法の延長線では取扱い得ない複雑な現象をどう解明し，理解するのかを主眼としているのである．従って，これには，多次元的な，ダイナミックな，可撓的な，探索的な考え方（The Way of Thinking）が重要な研究態度となる．

4．三つの相をバランスよく活用

データの科学には当然三つの相がある．一つはデータをどう計画してとるか（design for data という），どうデータを具体的に集めるか（collection of data という），データに対する解析（analysis on data という）である．大事なことはこの三つの相において一貫した考え方―データによる現象の解明・理解ということ―が貫流していなければならないことである．こうすることによって目的にふさわしく且つ妥当性ある方法が三つの相において，バランスをとって活用されることになる．データの科学の研究戦略の一つの例が「第1図」である．

一般的に言って現象は多様である．これは諸現象のフォーミュレーション，調査・実験の計画によって明らかにされよう．これが design for data の段階である．こうした考え方に基づいて実際にデータがとられることになる．得られたデータは計画と実施のあり方によって，その性格が評価されることになる．これが collection of data の段階であり，これらはデータの分析のあり方までも念頭に入れて行われる必要がある．

こうしたデータは多次元的であり，しばしば時系列データとして表現されることになる．得られたデータをどう眺めても，あまりにも複雑で，見通しが悪く，現象理解が難しい．そこで，多次元分析の諸方法，統計数理の諸

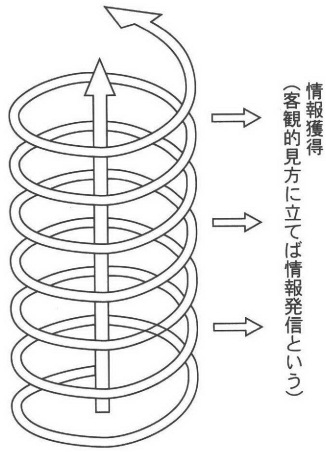


第1図　データの科学の研究戦略

方法，分類の方法によって単純化が行われ，集団構造が見出されてくる．単純な場合は平均値をとることによって見出されることもある．このようにして単純化，概念化が行われ，一応解明されたように見えてくる．

複雑な現象は，これで解ってしまう程，簡単ではない．こうした情報は，不完全で不満足ということも解ってくる．そこで，こうした構造発見・概念化を踏まえた上で，もう一度「個」に戻るのである．個を通しての集団構造の発見を摑んだ上で，もう一度個へ戻るのである．個の「平均値」からのズレ，構造からのハミダシを問題にすることになる．再び集団内における個の多様性をもとに考察を進めることになる．この新しい段階において再び単純化が講究されることになる．必要があるならば再び調査・実験が繰り返されることになる．このように循環的な研究が続いて行われることになる．単純化・概念化と多様化のダイナミックな相互交流が始まるのである．

一つの問題が解ったところで，新しい進んだ段階において，またいくつかの解明すべき問題が見出され，これを明らかにするための



第2図　上昇螺旋的研究の進展

研究が進むのである．データのデザイン→データの収集→データの解析→デザイン→収集→解析→デザイン…→という形でものごとが進むのである．行きつ戻りつ（progress and regress）しながら探索的に研究が進むということになる．つまり，上昇螺旋の形で研究が進められ，相互に知見の高まりを得ながら深く広く研究が進むことになる（第2図）．

以上は複雑な問題を取扱う場合の研究戦略の一例である．こうしてデータの科学は，理論や方法という方法論的成果ばかりではなく，そうした結果を編み出すために必要な諸過程に関係した諸種の方法論的成果をも包含するものである．つまり，データによる現象解明・理解に向けての全ての行為や結果をも含むものであり，きわめて広く常に膨張を続ける開集合的で且つしなやかなものであるというように考えられている．

以上が「データの科学」についての一応の根本理念である．

医学で言えば，医学は平均値であり，単純化のフェーズである．治療は，これを土台に多様性に対する多様化のフェーズである．この治療のフェーズから医学へのフェーズが円滑に行われることが望ましいのである．治療そのものの中から新しい医学が誕生するのが本道であると私は思う．基礎医学のみではなく治療のうちからこそ画期的な医学（やその方法）が生まれてきてよいのではないかと思う．

具体的にそれが三つの相でどのようになるか．社会（人間）現象に対するもの，医療に関係するものを念頭において説明してみよう．

5．分析のあり方まで見通して考えるデータの収集

データをどうとるか――．新しくデータをとる場合はどうなのか，既存のデータの場合はどのような計画の下にデータがとられているか，を第一に検討しなければならない．例えば次のようなものが挙げられる．

ア．標本調査の方法

どのようにユニヴァース（Universe）が決められているか，母集団（Population）はどのように定められているか，標本（Sample）はどのように抽出されているか．UPS の問題である．また，標本抽出計画の工夫はどのようにされているか．既存のデータであれば，UPS はどのようなものか，が講究されねばならない．これが，データによる現象の目的に対して妥当なものか，データの分析のあり方まで見通して考えることが大事である．既存のデータに対しても，この点を配慮しなくてはならない．

イ．実験の計画性

UPS をどのように考えるか．目的に妥当する「実験の計画法」を研究することは重要である．既存のものならば，その UPS はどうなのか．その実験の計画は，これから分析しようとする我々の目的に対してどのような意味を持つものなのか．

これらが基本であるが，検定論・推定論の呪詛を遁れれば，自由な考え方が様々に成立する．

このように挙げれば切りがないが，調査計画の段階で全体を見通して考えなければならないことは数多くある．ここが真剣に取り上げられなければ，いくらデータをいじくりまわしたとしても，本当にデータを解析したということにならないのである．

複雑な問題を取扱って，ある行為決定しようとする時，データの科学では，どのようにデータを計画して取ろうとするか，について一言触れておこう．

科学の方法は，比較的に単純で比較的に複雑な問題を取扱うところによく発達してきたもので，全く新しい複雑な問題を取扱うのは不得手なのである．ここを取扱う必要の要請に迫られている．そこで行われている方法は，伝統的なリジットな方法でデータを取り，一次元的な予測を行い，最適化に基づく検証を行うということが行われている．このために見当違いの議論が出てしまう．

我々は，「データの科学の考え方」により，フレクシブルに考え，「第１図」「第２図」に示した考え方を探索的に事を運びつつ，一つ解り一つ解らなくなりつつ，それを解くために更に進むという逐次近似的に進みながら情報を取り出し，最適化ではなく「危険の分散」という形で望ましい行為を評価しつつ―不明の点を考慮に入れ―情報を指し示すという形で進む事になるわけである．こうした立場で調査の実験が広く，高い立場で計画されるのである．

6．データの性格を客観把握・分析

ではデータはどう収集すべきか．計画に沿ってデータを収集するのだが，既存のデータならば，それがどう収集されているか，また，そのデータはどんな計画で作成され収集されたかを考え，その性格を客観的に把握しなければならない．この段階は単なる実務と考える人には「データの科学」を得る資格はない．収集にまつわる諸問題を，方法論的且つ理論的・具体的に研究しなければならないのである．例えば次のような問題が考えられる．

ア．調査・実験の歪みの評価

質問文による歪み，面接による歪み，調査員による歪み，観察による歪み，実験方法・条件による歪みなど，これに属し，データを取ろうとすれば必ず生ずる問題である．

イ．調査不能，実験における欠測値の評価

社会調査ならば調査不能，実験ならばデータの得られなかった条件分析などが必要である．

ウ．測定誤差の評価

エ．回答誤差・回答変動・不可避の測定値変動の評価

オ．関連する偏り，誤差の相殺されるような方法の工夫

これも挙げれば切りはないが，こうした相も方法論的に等閑視せずに厳格に取り上げることが大事である．

一方，データに対する分析についてはどうか――．得られたデータに相応しい分析の方法が採用されねばならない．特に，モデル構成のために不必要な数学的条件をおいたもの，精密な数学的条件の上に立つ方法を避け，数学的条件をなるべく課さない明解な方法が用いられることが望ましい．

ア．スケール理論，数量化，コレスポンデンス・アナリシス，多次元尺度分析法，探索的データ解析，カテゴリカルデータ解析，分類・クラスター化の諸方法など

これらが，データの性格に応じて適切に用

いられることが望ましい.

イ. 目的に応じた有用なデータ解析の方法の採用

ウ. 誤差, 偏りのあるデータの妥当な分析

エ. データの性格評価法とデータの質に応じた分析法

オ. 確率的回答に基づく分析方法

カ. 諸データ分析を活用しての探索的方法論の研究

キ. データの分類とそれらの構造把握の同時発見の方法

いずれにせよ, 三つの相が一貫した考えの下に,「データの科学の目標」に向かって取扱われることは妥当なことと思うのである. こうすると,「統計学もデータ解析も分類の諸方法」の発展も新しい方法に向かってくる. そうなると, それらの視点も高まり, 新しい地平線を見せてくれるものと思っている.

これまで十回に亘って異説統計学で「異端の話」をしてきた. 多くの方の習われた教科書統計学を否定する考え方であるのでついて行けないと思われたのではないかと思う. 書き足りないところもあるがあまり長くなると飽々するのでこのあたりで擱筆する. 書き足りないところは, 既存の書に発表してあるので (林, 行動計量学序説 前出, 数量化—理論と方法—朝倉書店, 1993) それを参照されたい. その項目をあげれば,

・因果関係を知ることはそんなに大事か
・モデルによる分析の虚妄
・データ解析としての多次元分析
・過程制御としての医学と治療
・QOL と治療のためのデータの科学

である.