データ解析からデータサイエンスへ
科学としてのデータを語る

企業のマーケティングは「経験と勘」から、「データ・情報」に基づく科学的展開が求められている。
基本となるのは「データ」。
しかしデータは単に存在するものではなく、質を評価し分析・活用することが大切だ。
生きたデータにするのはしっかりした手法がなければいけない。
基本となるデータをマーケティングなどの企業活動で生かすための扱い方などについて、
統計数理研究所の林知己夫氏に専門家の立場から「データ解析からデータサイエンスへ」とのテーマで、
“データを科学”してもらった。
ビジネスをちょっぴり離れ、学術的頭脳を刺激してもらう。(本誌編集グループ)

林　知己夫（統計数理研究所名誉教授）

難しい「本当の姿」の把握

データの重要性が叫ばれてから随分時がたった。マーケティングの世界でもデータ・ライブラリが整備され、多数のマーケティング調査が実施され、データの量は豊富になってきた。それをどう使いこなすかが問題になってきた。データはそのよって生じた過程によりさまざまな性格をもち、質もさまざまなものがある。これをそのまま通常のデータ解析にかけたところで妥当性のある情報を取り出すことはできるものではない。

さらにマーケティングにおいても、ただ一つの目的追求にのみ考えを絞って話を進めることが適切でない時代になってきた。幅広い、高い視点からものを見る必要が生じ、環境や社会的責任の問題をも念頭に入れることの重要性が認識されるに至った。

こうなると非常に複雑な問題をデータを通して理解しなければならなくなる。そのために、どのような調査・分析の戦略をたてればよいか――従来型の仮説・検証型のアプローチでは解決することができない――を考えなくてはならなくなってくる。データによって科学的に物事を処理する方法論が望まれてくる。ここに「データの科学」という考え方が不可欠のこととなってきたのである。

ここではマーケティングの問題を超えて、一般的に「データの科学」の姿を記述することにした――事象を限定することで却って大きな筋が見えにくくなるので――がマーケティングの諸問題との関係は深い。

データとは何なのか。データは「それらしい」定義はできても、データを作る専門家、データを分析することを専門にする人以外では、本当の姿を把握することは難しい。データは、ただそこに存在するのではなく、どのようにして発生したものか、あるいは作られたものか、という経緯によって性格を異にする。また、その質の良否も異なったものになっている。

統計学はデータ作成で始まる

データの質の評価なくして、データの妥当性ある使用は不可能なのである。当然のことのようであるが案外無視されている。Aという事象の出現率は70%と言っても、そのデータの発生・作成の方法によって意味は異なったものであるにも拘らず、その70%が独り歩きして事を面倒にしてしまうのである。発生・作成の方法を無視したデータは、むしろない方が良いことになる。玉石混淆のデータが氾濫している中で、これを活用するには、このデータの発生・作成方法に基づく真の評価と、それに応じた活用方法を考えることが第一に肝要なことになる。これが「データの科学」の第一歩なのである。

それでは「データの科学」(Data Science

)とは何なのか。統計学の現状からみてみよう。統計学はデータを作成することに始まり、データを分析し、結論を導くことを主眼としていた。特にデータの作成に関する、標本調査理論は、この分野で瞠目すべき考え方であり、方法であった。そこに用いられるユニヴァース(Universe、調査対象の集まり)、ポピュレーション(Population，母集団)、サンプル(Sample，標本)―三者を略称してＵＰＳという―の概念はデータの性格や質を「我々の目的」に対して評価する時の重要なポイントとなっている。

標本調査理論の考え方は、既存のデータの性格や質を評価する時の基準を与えることにもなったのである。これは統計学の最も大事な働きの一つである。しかし、標本調査理論は、平均値や総量の推定のみに焦点が当り―これによって重要な情報を与えることは言うまでもないが、複雑な現象はこれだけで十分なものではないのは当然のことである―様々な興味ある方法が生み出され、理論も精緻化してきた。つまり現象そのものの解明という点が呆けてきて、技におぼれるという観を呈してきた。現象解明のためには、堕落の道を歩み始めたことになる。

統計学の推定論・検定論もその当初においては、それなりの科学的妥当性をもっていた。しかし理論が進むと数学的精密化が行われ、現実から遊離した方向に理論が進んできた。実験計画法と言われる分野も全く同じ傾向を辿っている。総称して数理統計学という分野は、分化が進み、理論が高度化してくると、現象の解析という点から全く関係のないものになってきている。根源にあるＵＰＳの考え方すら無視された形になっている。つまり、進んだ数理統計学は、現象解明に関して、一般的に言えば無縁のものになりつつある。

こうしたことは、統計学だけの問題ではなく、多くの科学分野においても見られることで、「発展→分化→精密化」で活力を失うという形である。

複雑な現象を解明する手法が普及

データ解析の現状をみてみよう。データ解析は、数理統計学の方法ではデータの分析は不十分であるという点から出発し、単純な統計量統計学、統計的（形式的）推論という枠組みを離れ、データをいろいろ分析することにより、より有用な情報を取り出そうとする方法を考えるところから始まった。

1942年、東京帝国大学理学部卒。統計数理研究所所長を経て、現在、世論調査協会会長、興論科学協会会長、日本分類学会会長、統計数理研究所名誉教授など。標本調査、データ解析などに関する著書多数。

複雑な現象をも取扱うことを主眼として、さまざまな方法が工夫されてきた。質的・多次元的データの解析（数量化、コレスポンデンス・アナリシス）、分類・クラスター化の方法、グラフィカルな方法論などが活用されてきて、数理統計学で取扱わなかった問題を処理し、妥当な情報を与えてきた。データの活用・普及がこれによって大いに広まったのである。

ここまでは極めて順調であった。しかし、これに関連した理論を取扱う第二世代の研究者は、データ作成の意味が一向に解らず、やたらにデータを求め、これをいじり、理論を考えるようになった。データの手に入らぬ人々は、単純な構造を持つ人工データを用いて、既存の理論の性格を調べるようになった。人工データなら、その発生メカニズムを知って分析すれば一番良い結果が出るに決まっており、発生メカニズムの解らないという前提に立つ「データ解析の方法」が適切でないのに決まっている。このことすら理解しない研究者が理論の精密化を求め出した。

「そこはかとなき」情報を取り出す

データ解析の方法は発生メカニズム不明な対象を取扱い、探索的に情報を取り出すところに焦点があったのではなかったか。唯一無二の解を求めるのではなく、探索的にデータを彼方へ捻り、此方へ捻り、試行錯誤しつつ「そこはかとなき」情報を取り出しつつ進むところに特色があるのではないか。

これが可能になるためには、データの性格と質の検討から始めるべきでなのであるが、データ解析の理論は、この方面には進んできてはいないのである。「データ」らしきもの、人工データを土台とする理論の精密化、ブートストラップ法などによる推論化の方に進んでしまった。あるいは、便利なソフトの構築に力が注がれ出した。ソフト化は決して悪いことではなく、新しい方法論や方法・理論の研究のためにしなければならないこともある。しかし単なる便利なソフトの構築は、普及のため、他の分野への貢献のためには不可欠のことであるが、本来のデータ解析の方法の進展のためには必ずしも役に立たないのである。

現象の妥当な解析という目標を離れては、その分野の進展はない。ここにも、

分化→精密化による沈滞化が見られてきたのである。どうすれば切り抜けられるのであろうか。このためには、新しい概念を必要とする。かつて言われたライフ・サイエンス、ソフト・サイエンスという考え方も新しい概念であった。考え方であり、方向づけなのである。こうしたライフ・ソフトサイエンスで生まれてきた結果は、既存の目から見れば生物学の範囲に取り入れられるものであり、社会学、社会心理学、行動科学の範囲に属するものなのである。

既存の立場に立った人は、「目新しく論じるのは意味がない。生物学、行動科学で良いのである。香具師のようなことを言うな」と言ったものである。その通りのところもあるが、既存の枠の考え方からは、こうした新しい結果は生まれてこなかったのである。出てきたものは化け物ではないから、既存の範囲のものであるが、既存の目から生まれ得ることの出来なかったものが、新しい概念を作りあげた所に生まれてきたのである。私は、新しい方向や発展には、新しい概念が必要なものと思っている。

「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挙げてきた。

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、緊密な連繋をとることになり、第一回日仏セミナーが1987年に東京で行われた。それ以降、いわゆる「データ解析」(data analysis, analyses des donnees)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。

国際会議IFCS-96（神戸）で講演する筆者

1996年3月、国際分類学会連合(International Federation of classification Societies, IFCS)の国際会議IFCS-96が神戸で開催された折、本大会を特色付けるキャッチフレーズとしてData Science が用いられ、私が、What is Data Science ? - Fundamental Concepts and a Heuristic Example - （データの科学とは何か―根本理念と一つの説明例―）という講演をした。これをもとに「データの科学」の一応の説明を試みよう。

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。

3つの相をバランスよく活用

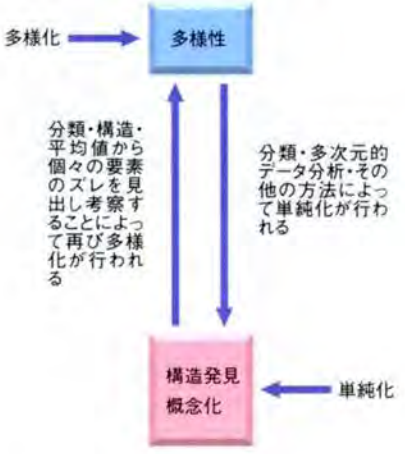
データの科学には当然３つの相がある。一つはデータをどう計画してとるか(design for data という)、どうデータを具体的に集めるか(collection of data という)、データに対する解析(analysis on data という)である。大事なことはこの三つの相において一貫した考え方―データによる現象の解明・理解ということ―が貫流していなければならないことである。こうすることによって目的にふさわしく且つ妥当性ある方法が三つの相において、バランスをとって活用されることになる。データの科学の研究戦略の一つの例が「図１」である。

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

図1●データの科学の研究戦略



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掴んだ上で、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（図２）。

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。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。

分析のあり方まで見通して考える

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。

ウ．クォータサンプリングのバイアスの評価

社会調査において、厳格なランダムサンプリングを用いることのできる国はごくわづかしかない。できる国であっても極めて費用が高いことに注目しなくてはならない。クォータ法を拒否すれば、社会調査のできる国は現実的にはなくなってしまう。そこでクォータサンプリングに関する調査計画法と実施法を詳細に調べ、クォータ法の質の評価ができるように考える必要がある。

エ．難しい複雑な問題を解くための調査計画をどう立てたら良いかに関する新しいシステマティックな考え方

オ．質問票の作成

ガットマンのファセット理論(Facet Theory)―これは仮説・検証の延長線上にあるもので、いわば理論主導型(theory - driven)の考え方である。

林の連鎖的比較調査分析法(Cultural Link Analysis, CLAと略称) ―これは、探索的方法を根幹とするもので、データを通して試行錯誤的に問題を解明しようとするデータ主導型 (data - driven) の行き方である。

いずれの方法も、質問票の作り方に関して一つのシステマティックな考えの下に組織的に行おうとするものである。前者は仮説をあますところなく包含する質問構成であり、分析を見通して作られたもので、うまく行けば見事な形が描けるが、予想外の発見のできる面白味はない。後者は、複雑で見通しの悪い国際比較調査を土台として考え出されたもので、諸国の同異の姿を歪みなく描きつつ探索するということが主眼であり、意表を衝く質問を加えることにより、未知の世界が

図2●上昇螺旋的研究の進展

開け、見えてくるという「わくわくする面白味」を期待できるという特色がある。

カ．質問における回答のとり方―選択肢法か自由回答方式か

それぞれ特色があるので、これをどう上手く使い、質問の狙いと回答の性格を明らかにするかが、その核心となる。

キ．投影法の社会調査における使用の工夫・恒常和法（おはじき法）による賛否の度合の測定

これによって心に思うことの深い意味を探ることが出来る。特に後者によって日本人の中間的回答の好みを明らかにすることができた。

ク．データの質・性格を評価するための調査デザイン

等質サンプルによる質問文の検討などがこれに入る。

ケ．ランダマイズ質問法の改善

回答者が調査員に回答を知られたくない場合、秘密投影法を用いる。もう少し自然にいわゆるrandomized response法というようなものを工夫することになるが、やはり、回答者に疑惑を与えるので、その方法を工夫する必要があるなど。

コ．国際比較調査における質問文の翻訳

国際比較は安易に行えば簡単なことである。しかし、何を比較しているのか不明な砂上楼閣のような議論もよく目にする。全く厳密に考えれば不可解なものである。何もしないよりはした方が遥かに豊かな知見になるという中間的な行き方を厳格に行うことが必要となる。このための方法論は既に述べたCLAであり、図1、図2に示した探索的な過程の上に立つ方法論である。

ここでは「データによって現象を理解しようとする」デーモンにとりつかれる事がなくては方法論は成り立たない。その一つとして翻訳の問題がある。原文→翻訳→再翻訳→再々翻訳のようなことの検討もその一つであり、自由回答による確かめ、再翻訳質問文による調査もその一つの方法となる。こうした細かい検討なくして国際比較ではデータによる現象解析は意味が薄くなる。

科学は「複雑な問題」が不得手

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「図1」「図2」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漸近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—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。

データの性格を客観把握・分析

ではデータはどう収集するべきか。計画に添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晰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．諸質問群の作成とそのコーディングとその総合

エ．誤差、偏りのあるデータの妥当な分析

オ．データの性格評価法とデータの質に応じた分析法

カ．確率的回答に基づく分析方法

キ．諸データ分析を活用しての探索的方法論の研究

ク．データの分類とそれらの構造把握の同時発見の方法

ケ．自由回答法の自動処理法

コ．確率的なアプローチ

サ．コンピュータ実験

これらも数え挙げれば限りなく出てくるが、その一例である。特に企業がマーケティングのために必要とするデータは、日常業務の過程で自然に蓄積されるものは少なく、計画的に収集しなければ存在し得ないものであり、収集と分析の諸問題は重要である。

勝手な分析は禍根を残す

既存のデータの活用であっても、データの収集のあり方を十分検討した上で、分析の方法を活用することが肝要である。既存のデータ・ライブラリをつくり、数列、カテゴリー列と見放して勝手に分析方法を施しても、どれほどのものか。禍根を残さねば幸である。ライブラリの使用も、データ分析法にのみ目をむけるのではなく、「データの科学」の立場からデータによる本当の現象理解に基づいて責任ある提言を行って欲しいものである。つまり、データに対する、そしてデータによる現象理解への情熱と感動が極めて根源的なものとなる。

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」に向って取扱われることは妥当なことと思うのである。研究者の立場から言えば、いつも全てにわたって研究することは出来ないので、個別の問題を取り上げて深めて行くわけであるが、データの科学の理念を中心として考え、最も重要と思う問題を取り上げて研究を進めて欲しいものである。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。