統計十話

第9話　探索的にデータを取扱うことの大切さ（No. 1）

—データ・マイニングとデータの科学—

林　知己夫

文部省統計数理研究所

これまで統計に関する話をまとめてきたが，いわば異説医学統計というべきものであった．つまり，医学教育・医学研究において正当的な方法として確立されているものを根底から覆すような考え方である．これが読者諸氏に受け入れられないことは重々承知している．糠に釘の感じである．間違ったことでも権威化されてしまっているものを打ち壊すのは大変なことである．頂門の一針という意味で書いてきた．いささかなりとも既存の医学統計に疑問をもっている方々への後ろ楯になればよい，この思いである．医学統計として確立されていることもなにがしかの貢献はしていたことを認めるのは吝ではないが，こうした数理統計の考え方そのものが医学研究そのものをその方向性までも含めて歪めていることも強く感じているものである．

二重盲検法によって利くと認定されたものを闇雲に使って治療するという考え方が正しいという発想を考えてみよう．私はこれはおかしいと考えている．全く少数の二重盲検法の単薬一定量の投与による結果は何を意味するのか．ごくわずかの情報しか与えていないことが無視されている．個体差，病態の多様な生きた人間の治療という観点に立てば，首をかしげたくなるのが常識である．この健全な常識をもちえなくなったところに医学統計教育の欠陥が見出される．二重盲検法の考え方は，製薬上の一つの目安であって，医療においてはこの情報の下に，いかようにして薬を利かせるか——個体差・病態と諸薬の量質を踏まえてのダイナミックス，栄養のあり方や心のあり方，QOL との総合処置の上に立って——という発想に立つべきで，このプロセスに有効な方法が，探索的立場に立つデータの科学なのである．私はこれを治療の科学化という形でまとめている（林：行動計量学序説．朝倉書店，1993）．

医学統計の発想の転換が，少なくともガン医学のまたガン治療の進展につながるものと思っている．硬直した医学統計の考え方がそれを阻害していると門外漢からはみえるのである．ここでは，データの科学とは何かというところを2回にわたって説明してみよう．

その前にこのごろ，データの科学の世界にデータ・マイニング（data mining）ということがいわれているのでふれておこう．これは，データのなかに宝を見出すということになろうかと思う．データ発掘，見すごしていることをデータ化し，こうして得られた諸データのなかに宝ともいうべき情報を見出そうというものである．日常の行為（医療行為）のなかに，データ化して有効なものがあるという観点に立つことである．経験しているが看過されているなかに宝があり，この情報を他の情報（他の人々の情報を含む）とともに集め

探索するところに宝の情報が出るという考え方である．他の人々の情報は発表された論文のみではなく，さらに幅広いものを考えてよい．つまり他の人々のデータ化された経験も含まれるのである．いわば，「机の抽出しのなかに何気なくしまい込んでいたもののなかに隠された情報がある」ということである．

さて，データの科学について，かつて「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」として日経ムックに書いたことがあるが，専門の違いから読まれたことはないと思われるので，必要なところをそれから抜粋してみよう．

1．はじめに

データの重要性が叫ばれてから随分時がたった．データはその拠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．

医学(治療)においても幅広い，高い視点からものをみる必要が生じ，人間・社会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――従来型の仮設・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．

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%といっても，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かかわらず，その70%が独り歩きして，こと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，むしろないほうがよいことになる．玉石混淆のデータが氾濫しているなかで，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はじまり，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「われわれの目的」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．これは統計学のもっとも大事な働きの一つである．

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．実験計画法とい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，一般的にいえば無縁のものになりつつある．

次に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はじまった．

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，多次元尺度解析など），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．

ここまではきわめて順調であった．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，単純な構造をも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よ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．

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はじめるべきなのであるが，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ほうに進んでしまった．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．

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．かつていわれたライフ・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った．

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，行動科学でよいのである．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うることのできなかったものが，新しい概念を作りあげ

たところ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．

「幾何学には王道はない」正にそのとおりである．従来の枠にとらわれない――しかしとらわれたがゆえにその知識のポテンシャルは高まったのであるが――ことが大事である．古い言葉であるが，「格に入って格を出ず」，その基盤として，素直にものを見ることである．