保健の科学　第42巻　2000　別刷

特集 新世紀の保健問題とコンピュータ

データの分析方法と結果の表現方法の変革はあるか？

高木　廣文

はじめに

今後十年ほどのデータ解析についての分析方法と，結果の表現方法で起こる可能性のある変革について考えたとき，あまりに漠然としており，どこから始めたらよいのか，とまどってしまう．しかし，自分自身のデータ解析に関する過去から現在までをまとめ，そこから将来について考えてみるのもおもしろいだろう．

将来予測というのは，通常は当たらないのが普通であり，あまり多くを期待されない方がよい．とはいえ，全くの見当違いというのも，後日恥ずかしい思いをするので，ある程度は論理的に考えをまとめていかねばならないだろう．この点から，データ解析に関して，自分自身が学生であった頃から，どのようなことを行なってきたのか，保健領域で必要とされる分析方法とは何か，そしてその分析結果の表現方法には何が望まれるのかを考えて行きたいと思う．そのような過程を通して，現在でのデータ解析での問題点や今後望まれる，また実現して欲しいことが浮き出てくれば何よりである．

筆者：新潟大学医学部保健学科

1．温故知新

およそ四半世紀前，初めて本物のデータの解析をした．私が卒業論文の指導を受けた教室が，厚生省特定疾患の全国疫学調査の集計を担当していた．そのため，そのデータを使用して論文を作成することになった．その時のおもな解析方法は，平均値や分散などの基礎統計学，クロス集計，そして因子分析であった．今でも似たような方法で解析している論文は結構ある．25年以上前でもそんな方法を使っていたのかと思う人もいるかもしれない．

現在と大きく違う点は，データ解析のためには，大型計算機を使用し，解析のためのデータはIBMカードという一万円札程度の大きさの長方形のカードに，80カラムずつキーパンチ（機械でデータを打ち込んでカードに穴をあける）しなければならなかったことである．当然，分析のための統計学的な手法もすべて自分でプログラムを書き，同様にIBMカードにパンチしなければならなかった．

2年後に修士論文を作成するときも，状況は似たり寄ったりで，やはりプログラムは自分で作成しなければならなかった．原稿も手書きで書くか，和文タイプライターを自分で打つかする必要があった．表を作成するにはIBMのタイプライターを使って行なった．ゴルフボール型の活字の部分

を適宜取り換えて表を作るのが，何か新しい感じがした．

それから3年後の博士論文の作成では，若干事情が変わっていた．すでにその3年の間に，統計学パッケージとしてSPSSが大型計算機センターに導入されており，比較的自由に使うことができた．とくに，クロス集計や基礎統計の計算，多変量解析の手法の中でも因子分析や判別分析では重宝した．しかし，多重ロジスティックモデルなどはまだSPSSには入っておらず，やはり自作する必要があった．また論文作成のための日本語ワープロはまだできていなかった．同様に，表はIBMのタイプライターを用いて作成していたし，図は大型計算機から出力されるものを縮小コピーなどをしたものか，ロットリングなどで線を引き，レタリングセットなどを使って，自分のセンスとテクニックで自作するのが普通であった．要約すると，20年前のデータ解析では，

①データ解析は大型計算機で計算．
②統計学パッケージは普及の初期であり，まだ自作のプログラムが必要．
③論文作成は手書きか，タイプライターを使用．
④表はタイプライターで，図は手書きか大型計算機の出力の切り張り．

というのが一般的（最先端？）であった．

それから約10年後，大学などの教員は個人でパソコンを保有しており，日本語ワープロ（英文ワープロも同様）ソフトも一般に普及していた．また，データ解析のためのソフトはすでに完成しており，われわれが作成したHALBAU(はるぼう)では，多変量解析の多くの手法が簡単に使用できるようになっていた．表は，HALBAUなどの分析結果をファイルに保存し，それをワープロソフトで読みとり，きれいに整えるということが普通に行なえるようになった．図に関しても，多くの作図用のソフトが開発され，特殊な図でなければ比較的簡単に作れるようになった．ただし，プリンターの多くはカラーに対応していなかったし，パソコンもデスクトップ型が主流であり，ノートパソコンもカラー液晶ではなかった．要約すると，

①パソコン（デスクトップ型）で計算．
②統計学ソフトの使用．特別な場合以外は，自作の必要はない．
③ワープロソフトで論文，表を作成．
④図はパソコンのソフトで作成．ただし，出力は白黒．

というのが，少なくとも私の状況であった．

現在は，どうかというと10年前とあまり変わっていないが，デスクトップのパソコンは，カラー液晶画面のノート型パソコンになった．これは私の趣味も関係しており，当然だが今でもデスクトップの使用者は大勢いる．統計解析用のソフトであるHALBAUは今も健在であり，変わったことと言えば，パソコンの基本システムがMS-DOSからWindowsに変化したことに対応して，Windows仕様になった点である．ワープロソフトや作図用ソフトなども，そのような変化に対応して変化しているが，最も変わった点は，文字の種類や大きさの選択がかなり自由にできる点と，カラープリンターの高性能化に伴い，色を比較的自由に使うことができるようになった点であろう．要約すると，

①パソコン(ノート型カラー液晶画面)で計算．
②統計学ソフトの使用．特別な場合以外は，自作の必要はない．
③ワープロソフトで論文，表を作成．
④図はパソコンのソフトで作成．出力はカラーも可能．

ということになる．

結局，20年前と10年前では，データ解析の環境に大きな変化が見られたが，その後の10年間では，それほど大きな違いは見られないことになる．実際には，より多くの人たちが，多変量解析などの方法を手軽に使用できるようになり，より簡単にデータ解析を行なうようになったという点では大きな変化があったといえるだろう．

この10年間で，分析やワープロとして使用する計算機は，高機能，小型化，軽量化が加速されていることである．一方，分析，ワープロ，作図などのためのソフトは，高機能になってきている

が，必要な操作を覚えるのが必ずしも容易ではなくなってきている．

このようなデータ解析に関する環境の変化を踏まえて，10年先の変化を考えてみよう．

2．分析方法の変革はあるのか

データを解析するためのツールは，ここまでに述べたように，小型化かつ高性能化されてきた．一方，分析方法の変化はどうであっただろうか．20～25年前と比べると，使用している方法のうち主要なものにそれほど変わりはない．そうは言っても，細かく見ていくとそれなりに違いもあるようである．

全く変わっていない方法として，

①基礎統計量の計算．

②クロス集計/相関係数とその検定．

③2群の母平均値の差の検定（t検定）．

④分散分析など．

以下きりがないが，データ解析の基本的な方法は，まったくその価値も重要度もそのままである．

ほとんど変わっていないが，統計学ソフトが手軽に使えるようになったために，

⑤主成分分析，因子分析，重回帰分析，判別分析，クラスター分析など．

従来から使用されていたが，より手軽に使用されるようになった．

20年前はほとんど使用されていなかったが，ここ10年ほどかなり使用頻度が増加した方法として，

⑥多重ロジスティックモデル，Coxの比例ハザードモデル．

があると考えられる．この2つの方法は，現在では疫学研究の標準的な解析方法のひとつになっている．

さらに，行動科学的な研究では，

⑦共分散構造分析

も使用されるようになった．この理由は，この方法が「因果モデル」を解析できるという誤解に基づく，過剰な期待がその使用の背景にあるように思われる．実際には，共分散構造分析は，研究モデルのグラフ化という点にその利点があると筆者には考えられるのだが．

この他に，以前から言われている割には大して進展していない方法に，

⑧エキスパートシステム

がある．これは，例えば因子分析を行なう場合，その方法についてほとんど知識のない人でも解析を十分行なえるように，データの入力から解析，そして解析結果の解釈までを，コンピュータが相談や指示をし，サポートするというシステムである．残念ながら，完全にうまくいくシステムを作ったという情報は，今のところない．

以上のような点から考えると，研究に必ず必要とされる基本的な解析方法が変わるようなことはあり得ないだろう．すなわち，①～⑥の方法は，保健学分野でのデータ解析で，今後もその重要性に変化があるとは考えられないからである．調査対象者の背景の記述や，因子の探索，特定の事象の予測や判別，ある疾患に対するリスクの検討についての必要性が，今後10年間程度の短期間に大幅に変わるとは到底考えられないからである．

それでは，何も変わらないかというと必ずしもそうとはいえない．統計学専門の学会誌などでは，各種の方法が提案されているからである．多くの手法は，極めて特別な場合にしか有効でないか，それまでの方法の一部修正というものであり，これを使用すればそれで何もかも済むというものではない．多くの方法が提案され，そのうち運良く統計学ソフトに採用された方法だけが，生き残っていくはずである．

結局，分析方法自体の大幅な変革はほとんど期待できないが，そのツールの変革は大いに期待できる．すなわち，パソコンの更なる小型化と携帯電話などの結合による「データ解析のモバイル化」である．これは，データ解析のインターネット化の延長といえばよい状況が，一般に普及した状態と考えればよいだろう．データ解析をパソコンで行なうだけでなく，何処にいても適当なホームページにアクセスして，その資源を利用して解析

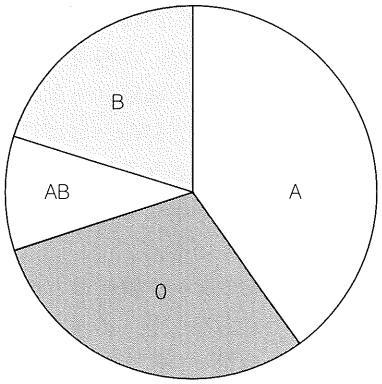

図１　血液型の円グラフ

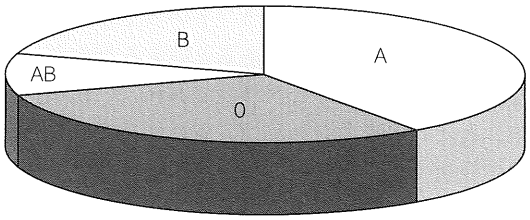

図２　血液型の立体円グラフ

を行なうのが普通になっているだろう．

３．何が結果の表現で重要か

分析結果の提示の仕方で何が変わったかというと，

①ワープロ（ソフト）の普及で，学会発表用の抄録，表，図が手書きでなくなり，きれいになった．

ということが，過去20年間ほどで最も大きな変化であろう．また，カラー印刷が手軽にできるようになったため，

②学会でのポスターがカラー印刷となり，分かりやすい表示が増えた．

これは，作図用のソフトの普及も大きな力となっている．さて，分析結果の提示の方法で何が重要なのか，ここで一度考えてみよう．

統計的な解析結果は，通常は数値によって与えられることが多い．したがって，その結果を正確に提示するには，表形式で表すのが最も分かりやすいといえる．パソコンの普及に伴うワープロソフト類の普及により，例えば，因子分析や重回帰分析などの結果を簡単に，かつ美しく作成することが容易になった．学術的な研究での表の使用は，今後も現在同様に必要であり，とくに表現方法に工夫をする余地はほとんど残されていない．たまに，どうしてこんなわかりにくい表を作るのかと思ってしまうようなものにお目にかかることもあるが，それは個人のセンスの問題であり，ここではこの点については考慮しない．

図に関しては，表に比べると，その表現方法は多様化しているといえる．ひとつには，白黒以外の色の使用が可能になったことと，立体的なグラフ表示が増えたことである．今後も，ポスターなどではカラフルで立体的なグラフが増えると考えられるが，実はここに大きな問題がある．

科学的な研究では，表は正確な分析結果などを示すために不可欠であるが，同様に図は分析結果を「ある程度正確に」分かりやすく示すために用いられている．問題とは，立体的なグラフは何となく新しく格好いい感じがするが，必ずしも正確な印象を見る者に与えるとは限らない点である．簡単な例として，**図１**と**図２**を比べてみよう（本当はカラー印刷が望ましいのだが）．この２つの図は同じデータを元に，一方は従来からの平面的な円グラフで血液型の分布を示したものであり，他方は現在流行（？）の立体的なグラフである．**図１**は正しくその面積（もしくは角度）で各血液型の割合を示している．一方，**図２**も立体的に作成したらそう見えるはずの各血液型の割合を表してはいる．しかし，AB型は実際よりも少なく感じられるし，O型とB型，A型を比べたとき，直観的に正しくその大きさの順を把握することはそれほど簡単ではないだろう．

この簡単な例から分かるように，見栄えの良さそうな方法が必ずしも，方法論的にみて良いとは限らないことが理解できるだろう．それではなぜ，立体的なグラフは正確な情報を与えないのだろう

か．その理由は，立体グラフといっても2次元平面に表示しているためであるといえるだろう．したがって，どのようなグラフであれ，それが正しく3次元的に表現されれば，極めて分析結果の理解に役立つだろう．例えば，因子分析の結果を3次元空間図で正確に表すことができれば，2次元平面での配置図に比べ，変数相互間の関係を把握するのが容易になるだろう．現在でも，立体散布図のようなグラフは存在するが，決して見やすいものではない．

3次元的に正しくグラフを表示するためには，バーチャルリアリティの技術が必要とされる．パソコンの超小型化とともに，現在の液晶ディスプレイが眼鏡タイプのものに移行していけば，グラフの立体表現は比較的近未来に実現されると考えてよいだろう．現在のような一方方向からの平面的立体グラフではなく，あらゆる方向からグラフを眺められるようになり，ごまかしによる表示ではなく，正確な情報を与えるためのツールになるだろう．

4．近未来のデータ解析

古いSF映画では，宇宙ロケットの中で使用しているコンピュータは，入力も出力も紙テープか出力はプリンター出力になっていたりする．さすがに現在の映画ではそんなことはないが，人間の想像力が如何にその時代の状況に縛られているかを，端的に示すものだと思う．

初めてNECのPC-8000に触れたとき，それまで使用していた東京大学の大型計算機とのあまりの落差に驚き，これでどうやって多変量解析が行なえるのか悩んだのが，およそ20年前である．初めて，ラップトップ型のパソコンを買ったのは12年前であり，持ち運ぶのは結構体力が必要であった．そのような状況下で常に考えていたことは，そのうちより軽量で画面もカラーのノート型のパソコンが開発されるはずであり，それで必要な統計計算や論文作成ができるはずであるということであった．もっとも，当時は「死ぬまでには」という程度の期間を想定していたのだが，10年もかからずに実現してしまった．このような思いを抱いていたせいか，とにかく私はノートパソコンが好きで，絶対デスクトップ型のパソコンは購入しない．HALWINの開発も，6万人以上ある生活習慣調査データの解析も，すべてB5判サイズのノートパソコンで行なっている．

それでは，近未来のデータ解析はどのような状況で行なわれるのであろうか．まず，そのためのツールとして，大容量（1テラバイト超）超小型（現在のMDプレイヤー程度）のパソコンの出現はほとんど確実であろう．当然，コードレスでインターネットに接続可能なタイプであり，必要なソフト類は全てインターネットからダウンロードするのが一般的になっているだろう．音声入力機能，手袋タイプの入力用センサー，眼鏡タイプのディスプレーが標準的な装備となっているだろう．

このような状況では，いつ何処でもデータ処理が可能であり，大学や特定の研究室でなくとも，高度な解析を行なうことが可能となり，得られた結果を即座にグラフ化することも簡単に行なえるようにソフトが開発されていることだろう．さらに，研究結果を展示するのにポスターなどを作成する必要はなく，自分のホームページ上に掲載するのが一般的な方法となっていよう．ホームページ作成のツールは現在すでに開発されているが，より簡単に，よりグラフィカルになるのが必然的な動向といえよう．

今のところ，統計学ソフトは満足に使えない，Excelでグラフがうまく描けない，インターネットなんかとんでもないし，ましてホームページを作るなど考えられないという方も心配は無用である．技術の発展の方向は，確実に人を楽にさせるように向かっている．正しいデータ解析と正しい結果のプレゼンテーションが，より容易にできるように，ハードのみならずソフトもやや遅れながら，現実化されていくと考えてよいだろう．