統計数理（1997）
第45巻 第1号 23-40

データ解析からデータ・サイエンスへ

——情報技術（統計ソフト・WWW・AI）との共生により統計知識を国民の知的共有財に——

成蹊大学* 新 村 秀 一

（1996年7月 受付）

1. はじめに

矢島・大隅（1977）により，海外の統計ソフトが日本に紹介されたのは1977年である．また，SPSSが日本の大学で使われ始めたのは1969年であり，筆者がSASを住商情報システム(株)（略して，SCS）に導入したのが1977年である．

それから約20年たった昨今，表1に示すように，統計ソフトに関する話題が再び増えてきている．

このことは，多くの統計分野の関係者が，「統計ソフトによる新しい局面を求めて，再び活動していかなければいけない」と考えている証拠でなかろうか．

その理由としては，次の点が考えられる．

- 統計ソフトが，大学の研究者にとどまらず，企業の実務にも使われ，利用経験の蓄積と問題点がある程度明らかになってきた．
- 情報技術の進歩により，大学や企業の汎用機で使われてきた高額な商用統計ソフトが，PC

表1． ここ2-3年の統計ソフトに関するイベント（主として筆者が関係したもの）．

1994年11月	オペレーションズ・リサーチ誌11月号で，「統計ソフト」に関する特集． 特集「統計ソフト」：オペレーションズ・リサーチ誌，第39巻11号（1994）．
1995年7月	大分大学で開かれた日本統計学会で，大橋先生（東大）の企画で統計ソフトに関するチュートリアルが開かれた． 1995年度統計学会チュートリアル・セミナー予稿集（1995）．
1995年10月	垂水先生(岡大)の推薦で筆者が大会会長を務めた第9回日本計算機統計学会シンポジュウムで，「日本における統計ソフトの過去・現在・未来」というテーマでシンポジュウムを開いた． 第9回日本計算機統計学会シンポジュウム（テーマ：日本における統計ソフトの過去・現在・未来）論文集（1995）．
1996年9月	幕張で開催される日本統計学会では，柴田先生（慶應）がオルガナイザーとして「統計をデータサイエンスとして発展させよう」という趣旨のセッションが行われた．日本統計学会第64回大会予稿集（1996）．
1996年11月	北海道で開催される第10回日本計算機統計学会シンポジュウムでも，田中先生（岡大）の企画で統計相談のセッションがある．

* 経済学部：〒180 東京都武蔵野市吉祥寺北町3-3-1．

で利用できるようになった．PC上の統計ソフトの特徴は，価格面では汎用機版に比べ比較的廉価であり，ユーザーインターフェイスがCUI (Character User Interface) からGUI (Graphical User Interface) になった点である．これにより，組織人としてしか利用できなかった商用統計ソフトが，無理をすれば個人が自宅でも利用できるようになった．さらにGUIにより，ユーザーが複数の統計ソフトを使い分けることも容易になった．

・インターネット特にWWWの技術により，個人や学会が，比較的容易に統計のノウハウや情報を国内や国外に発信できる環境が提供されてきている．

以上の理由から，これからは，統計知識をこれまで以上に国民の共有知識として普及し，大学の教育や研究に加え，産業界でも多用される時代を創設するための環境が整ったといえよう．本稿では，そのための目標や方法に関し提案したい．

2. 統計ソフトの世代分類とデータ・サイエンス

筆者は，統計ソフトを，表2のように3世代に分類することを提案したい．

2.1 第1世代：自作ソフト（統計理論が主役）の時代

第1世代は，自作ソフトの時代であった．個人が，統計アルゴリズムの他，コンピュータ言語やコンピュータ・システムを熟知してプログラムを作成し利用するという，職人芸や個人芸の時代であった．この時代は，統計理論が主役であり，理論を補足するのがその大きな役目であった．その意味で，「統計理論の時代」ともいえる．そして開発されたプログラムが，解説書の付録のリストで公開されたり，MTでソースコードが配布され，漸く企業で利用されもした．

2.2 第2世代：汎用統計ソフトによる普及（データ解析が主役）の時代

第2世代の汎用統計ソフトによる普及の時代は，第1世代の自作ソフトの時代と比べての話であるが，次の点が著しく改善された．

・ソフト資産の共有化により，自作していたのではとうてい利用できない多くの統計手法が，統合的に研究者のみならず一般ユーザーにも利用できるようになった．
・ソフトの保守から解放され，統計本来の仕事ができるようになった．
・解析結果の信頼性が向上し，プログラムのバグやアルゴリズムの検証に時間をとられなくなった．

表2．統計ソフトの世代分類．

第1世代	自作ソフトの時代
第2世代	汎用統計ソフトによる普及の時代
	ソフト資産の共有化
	保守からの解放
	信頼性の向上
	汎用機からWSへ
	企業の情報系と大学，研究機関での組織利用
第3世代	データ解析からデータサイエンスへ
	汎用統計ソフトとネットワークの共生の時代
	PCとGUIによる個人利用の促進
	インターネットとAIによる知識の共有化

・解析結果が利用者間の共通言語となり共有財産になった.
・大学で統計教育を受けていなくても，汎用統計ソフトを用いて実際のデータを解析することを通して，研究者以上に統計利用者人口が増えた.

すなわち，汎用統計ソフトによって，統計が大学の理論研究から，一部企業の実務で利用されるようになった点であろう．特に，医薬品業界，各種調査会社，新聞社，金融機関等である.
統計理論をふまえ，実用化が促進された「データ解析の時代」と呼べば良いであろう.

2.3　第3世代：データ解析からデータサイエンスへ

1985年以降，汎用統計ソフトのSASが汎用機やWSからダウンサイジングし，PCでも稼働するようになった．さらに，1990年代に入ってPCの価格/性能比が飛躍的に向上し，使いやすいOSのWindowsが普及し，SPSS for Windowsのような新しいGUI版の商用統計ソフトがレンタルから売りきりで販売されるようになった.
これにより，個人が自宅のPCでもデータ解析の仕事ができるようになった.
すなわち現在は，情報処理技術の発展により統計ソフトにもパラダイムシフトが起きている．それは，第1世代の「自作ソフト（統計理論）の時代」が完全に過去のものになり，第2世代の「汎用統計ソフトによる普及（データ解析）の時代」がまさに終わろうとしているのではなかろうか.
それでは，次にくる第3世代は何であろうか？
それは，汎用統計ソフトがインターネット等の他の情報化技術と共生し，企業や大学などの一部の統計愛好家に限られた統計の知識を，ホワイトカラー全般の知的共有財産としてさらに普及させることではなかろうか.
例えば，企業のライン管理職が表計算ソフトによる集計計算に加えて，統計ソフトで意思決定を日常的に行うような状況を一つの目標にすべきである.
そして大学では，もう一度統計教育を，統計ソフトや他の情報技術を取り入れた情報リテラシー教育として再編し，多くの学問の触覚や道標になることを目標にすべきであろう．これまで数学が，自然科学の共通言語であったように.

2.4　データ・サイエンス

ここで唐突に第3世代という分類を導入し，それを「データ・サイエンス」の時代とするのには問題があるかもしれない.
柴田（1996）は，第64回日本統計学会の「データ・サイエンス」に関するシンポジュウムで，

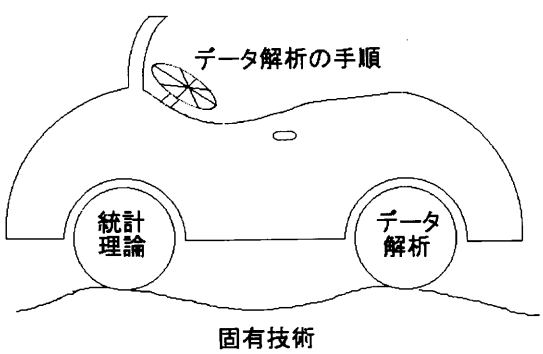


図1.　データ・サイエンスの概念図.

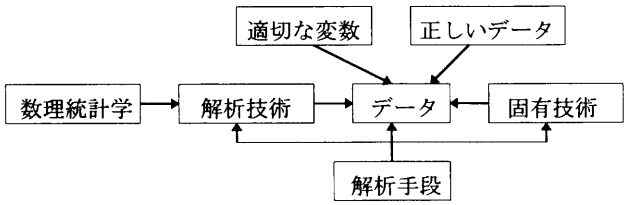


図2. データ・サイエンスの概念図（芳賀）.

データ・サイエンスに関する林（Hayashi (1996)）の見解を，本人は必ずしも全面的に同意しないと断った上で，次のように紹介している.

それによれば,「これまで独自の道を歩んできた統計学とデータ解析を統合する概念としてこの用語を用いている．また，データ・サイエンスとは『データによって示される実際現象を解析しようとするものである』と定義づけている.」とまとめている.

同シンポジュウムで発表とパネル・ディスカッション（大橋，大隅，垂水，新村）に参加した新村 (1996a) は，図1を例に示し，データ・サイエンスという概念はまだ定着していないが，とりあえず借用しても良いのではないかとの意見を述べた．実際のところ，これまではデータを扱うアプローチを「データ主義」と呼んできたが（新村 (1993)），データ・サイエンスあるいはデータの科学の方が呼び方として適切と思う.

すなわち，表2と対応して考えれば，統計理論を車の前輪として，その後の統計ソフトにより発展した後輪のデータ解析によって，実際のデータを運ぶ（解析）ことができるようになった．そして，ハンドル捌きとしての解析手順の方法論と，固有技術の理解が今後重要になってくる．固有技術の裏付けのないデータ解析は，例えてみれば空を飛ぶ車のようなものであろう.

これに対して，芳賀（1996）から，図2の概念図を示された．データを中心として，数理統計学（統計理論），解析手段（データ解析），解析技術（解析手順の方法論）と固有技術の関係が矢印で示されており，図1より役割が明確である.

しかし，冷静に反省してみると，これまでのデータ解析では，個別データの解析に目をとられ，データそのものを科学的にそして普遍性をもって研究してこなかったし，その環境にもなかったわけである.

そこで漸く，椿（1996）や大森（1996）によるデータ管理と品質に関する発表や，垂水・山本（1996）によるデータ辞書などの，データに関した一般化された研究報告もみられるようになった.

すなわち，今後新たに検討すべきは，データそのものでありデータの解析技術である．これらをより一般化・普遍化し，これまで以上に多くのユーザーが，統計ソフトを多くの側面で，誤用せず，多用できるように時代をリードすべきと考える.

以上の意味で,「データ・サイエンスの時代」と呼んでも良いのではないかと考える.

3. 統計ソフトの研究

第2世代では，表3に示す様々な海外の商用統計ソフトが開発された.

統計ソフトに関しては，商用と学術的なものがあるが，ここでは筆者の知り得ている商用ソフトに限って議論を進めたい．日科技連の JUSE シリーズは，教育やセミナーと連携して品質管理の分野に関して大きな成果をあげている．しかし，品質管理の実務経験や JUSE の使用経験がないので割愛した.

表3.　商用統計ソフトの特徴.

SAS	統計システムの開発言語，プログラム言語と統計ソフトの両面で利用
SPSS, VisualStat (Statistica) StatPartner, Jump	データ解析
S	アルゴリズム開発と教育
表計算	簡単なアルゴリズム教育，データ保管と配布
Mathematica, Mathcad	統計実験
TSP	経済学部での利用

3.1　統計ソフトの選択・利用基準

商用統計ソフトは，ブラック・ボックスであることと，ソフト一般のもつ性質であるホット(熱烈なユーザーになりやすい)な点から，群盲象をなぜ雲をつかむ感がする．このため，商用統計ソフトのユーザーの間で，お互いにコミュニケーションができず，比較評価が難しい点があった．

例えば，大隅・秋山（1988）は，PC98で稼動する国産の統計ソフトに関して，基礎統計量と回帰分析に限って比較評価しているが，それだけでも大変であったと考える．

しかし最近のWindows版の商用統計ソフトでは，操作性がある程度統一され，筆者を例にとればやっと複数の商用統計ソフトを使い分けることができるようになった．

私自身は，数年前まではSASユーザーであったが，組織人としてしか利用できないこと，PC用の廉価な統計ソフトが現れたことから，SPSS for WindowsとVisualStat (Statistica) の個人ユーザーになった．また，SpeakeasyやSAS/IMLの使用経験から，Sを行列演算のできる同種のソフトとしてとらえている．表計算，Mathematica, Mathcadの利用に関しては，本当の初心者である．

以上の限られた経験から，これらソフトを大まかに特徴づければ次のようになろう．個々の手法や機能に関する具体的な評価は，今後の検討課題である．

(1)　SAS

SASは，統計アプリケーション・システムの開発言語と考えている．筆者は，1996年3月まで企業にいたが，そこでは統計アプリケーション・システムの開発言語として利用してきた．筆者のいた企業では，SASを用いて製薬企業，金融機関，電力・ガスなどの公共事業，リクルート分野などの統計システムの開発と受託計算やコンサルテーションを行ってきた．

4月以降は大学に移籍したが，その2年前から非常勤講師として大学の情報処理教育やデータ解析を担当するようになった．そこでは，SASを用いているが，他のWindows版の統計ソフトが飛躍的に使いやすくなったため，相対的にSASの統計ソフトとしての優位性はなくなったと考えている．このため，SASを他のWindows版の商用統計ソフトと差別化するとすれば，統計の他にプログラミング言語として併用して教えることだと考える．

すなわち，統計が重要なアプリケーションの開発にはSASが適している．しかし，単にデータ解析として使用する場合，豊富な統計手法は捨てがたいが，価格やユーザーインターフェイスの点で，以下に述べるWindows版の商用統計ソフトに対して，優位性がなくなったと言えよう．

筆者が，以下で統計ソフトのパラダイム・シフトを唱えるのも，汎用機上のSASユーザーだったことが大きく影響しているのかもしれない．

(2)　統計ソフトのパラダイム・シフト

筆者がSASを利用し始めた時点で，世界中でSPSSは2000ユーザー以上，SASは500ユー

ザー程度であった．さらに，SPSSには日本語による解説書や国立大学大型計算センターという強力なユーザーがあった．

しかし，SASスーパーバイザー（ユーザー開発プログラムの登録機能を含む）と行列言語というような他の統計ソフトにはみられない機能を評価して，当初導入を考えていたSPSSをやめ，1977年にSCSへのSASの導入に踏み切った．そして，仕事として前にも述べたアプリケーションの開発や受託計算を行ってきた．一方SASは，個人的には，ライフ・ワークの一つとしてのデータ解析の家庭教師でもあった．

統計ソフト以外にもORや数値計算などのソフトウエアの普及に努めてきた経験から，「汎用機で成功したソフトウエアは，技術的な問題よりも過去の成功メカニズムが壁になり，ダウンサイジングに失敗する」との持論をもつにいたった．例えば，利益の大きい汎用機でのレンタル商売を変更できないのは，何もIBMだけでなくSASも同様である．

また本稿ではこれ以上触れないが，「統計，OR，数値計算・数式処理，AIの技術系ソフトは，最終的には独立系のソフトウエア会社による世界市場をターゲットにしたものだけが残る」という信念もここ10年来のものである．

一方ユーザーの立場から言えば，統計ソフト等の意思決定に用いるソフトは，個人でも利用したいものである．30歳代初期の頃は，新村・三宅（1983）と新村（1996b）を作成するため会社のIBMの汎用機を半日利用し，SASの回帰分析の総当たり法(52万モデル＊3回)を行った．しかし，このような個人研究は，黙認されたとはいえ厳密に言えば会社資産の無断使用である．そこで，高森・新村（1987）の執筆に際しては，無断使用していない証とするため，会社に導入していないPC/SASをわざわざSASソフトウエア(株)で借りるようなこともした．

筆者は，統計やORは人間の意思決定に役立つ学問であり，会話型ソフトウエアがそれを実現する要であると考えている（新村（1993））．その点で，企業人としてでなく個人としても自由に会話的にこれらを利用できることを夢見ていた．

1985年以降，SASやSPSSがPCにも移植された．しかし，汎用機と変わらないCUI版のPC/SASは，PCの能力不足もあって，魅力に乏しかった．レンタル制のSASと異なり売りきりになったSPSS/PC版も，SASに対する従来からの愛着と惰性もあり，興味がわかなかった．

しかし，Windows版のワープロを体験し，SPSS for Windowsも同じ操作法で利用できるので，高森・新村（1987）で用いた「学生の生活実態調査データ」（40個体＊8変数）を用いて新村(1994b)を執筆した．すなわち，SASコマンドやSPSSのシンタックスという独自な言語から解放され，Windowsの共通な操作法で利用できるようになった．ここ数年で筆者のように複数の統計ソフトを使い分けるユーザーも増えてきているようだ．

この他，

- 高品質のグラフ出力が標準になった．
- ワープロなどの他のWindowsソフトと連動して使える．
- PCが10年前の大型汎用機と同等あるいはそれ以上の能力を持つようになった．ユーザーインターフェイスの使いやすさを考えれば，それ以上の環境を個人でも容易に手に入れることができるようになった．

すなわち，商用統計ソフトが組織から個人利用へ拡大されるというパラダイム・シフトが起きた．

(3) Windows版の商用統計ソフト

データ解析の教育や実務には，SPSS for Windows, VisualStat (Statistica)（関田（1995）），StatPartner（吉田・石橋（1995））等のGUIベースのものが，価格や操作性や会話性やグラフによる理解のしやすさからよいだろう．また，個人でも利用できるようになった点が重要だ．

この他，SAS の Jump や Compstat '96 を目標に開発された NAG (SCS が代理店) の Genstat for Windows もあるが，国内の販売体制に問題があるようだ．

しかし，大学の統計教育には実績のある汎用機や WS 上の SAS がまだ多く使われており，PC での統計教育はこれからの課題であろう．

(4)　行列言語

第 1 世代の自作の時代には，多くの技術系のプログラムが Fortran で記述された．Fortran のようなプログラム言語は，処理単位がスカラーである．そこで配列や行列を処理単位とする言語 (行列言語あるいは中間言語) が現れた．例えば，プログラム言語としての APL や，数値計算パッケージの Speakeasy である．数値計算では，行列は重要であるが単発に近い固有値計算や行列演算が主体であったためか，Fortran の強力な代替になり得なかった．

しかし，統計分野では，統計アルゴリズムの多くは行列で記述できるが，配列処理や出力の編集というプログラミングの機能も必要であり，Fortran のようなプログラム言語の機能と配列・行列を処理単位とする機能を併せ持った S や SAS/IML の特徴が注目される点であろう．

すなわち，統計アルゴリズムの開発や教育に重点があれば，汎用統計ソフトに比べて，S や SAS/IML の利用が考えられる．

(5)　表計算ソフト

表計算ソフトは，企業では，業績の集計や営業管理情報の集計などに多く用いられている．統計ソフトの利用をこのレベルに近づけたいというのが，筆者の主張である．

一方大学では，価格の利点から，表計算ソフトで統計教育を行う向きもある．特に，経済などの人文科学系では顕著である．

単に汎用統計ソフトの価格が高価なため汎用統計ソフトを使わないのか，表計算ソフトで統計教育を行うことが汎用統計ソフト以上のメリットが得られるのかは，今後の検討課題であろう．

あるいは，「大学や企業の現状は，高度な統計知識が一般化していないので，表計算ソフトに含まれる統計量の程度で十分」という意見も多いが，本当にそれでよいのか真剣に検討する必要があろう．

筆者は，汎用統計ソフトではできない簡単な基礎統計（平均・分散など）や相関の計算方法の教育に用いている．しかし，それ以上の検討を行っていないし，表計算ソフトによる統計の解説書も検討していないが，データ解析をスムーズに行ったり，データ・サイエンスの主役になり得ないと考えている．

今後は，表計算ソフトに限らず，各種の統計ソフトによる統計教育に及ぼす影響を比較評価すべきであろう．

一方，複数の統計ソフトが存在する以上，データの互換性が問題になる．テキスト形式が第 1 に考えられるが，表計算ソフトを介在させることで異なった統計ソフト間のデータ交換やデータ配布の媒体に用いることが考えられる．

(6)　数値計算・数式処理ソフト

Mathematica や Mathcad は，汎用統計ソフトの不得意な統計実験（シミュレーション）に適しているが，データ解析には使えない．

Mathematica ユーザー会で，Mathematica を統計ソフトとして普及させようという意図のセッションが開かれた．新村 (1994c) は，「学生の生活実態調査データ」を用いて「データ解析には使えないが，確率統計の理論を教えるには使える」との主張を述べている．

一方，汎用統計ソフトは，乱数は発生できるが確率分布関数を重ね書きしたりするのが容易でないなど，確率実験に用いるには今後改良の余地があろう．

(7) そ の 他

特殊な例として，経済学部ではTSPなどが用いられているし，多くの大学ではフリーソフトや廉価なソフトの利用も多いようだ．

これらは，当該分野の必要とされる機能を十分に満たしていれば，あえて汎用統計ソフトを使う必要はないが，他ソフトとの十分な比較検討がなされていないようだ．しかも，どの手法をどの学部で教えるかといった議論も経済学部にかぎらず不十分のように思う．

3.2 統計ソフトを見る視点

統計ソフトを評価する視点として，表4の項目が考えられる．ただし，個人利用，企業や研究組織での組織人としての利用，大学などの教育利用などの目的によって意味合いが異なってくる．ここでは，主として大学などの教育目的として考えてみる．

(1) 価　　格

まず，大学での利用を考えるとき，価格が問題になる．

SASやSPSSやSが教育に使われている大学は，比較的恵まれた環境にあるといえよう．多くの場合は，価格の安い表計算ソフトか，フリーソフトや自作ソフトでの教育，あるいはあきらめている場合もあろう．

上田（1996）は，統計教育ソフトの問題点について報告しているが，統計ソフトを用いた情報処理教育の実態と併せてさらに検討していく必要があろう．

一方，第2世代において先駆的な企業では，費用対効果が重要で，価格が特に重要な問題にならなかった．それよりも統計の有用性を社内で説得できる人材のありなしに導入が影響されることも多かったようだ．

しかし，今以上の統計利用人口の拡大を組織と個人利用で目指すなら，他のPC用のソフトウエアに比べて，割高なのも事実である．ソフトウエアは，製造コストが限りなくゼロに近づくので，ユーザー数が価格の決定に大きな割合を占める．普及のため低価格にするか，普及すれば低価格になるか，鶏と卵のジレンマがあるのも事実である．

価格問題に対しては，現状では商用ソフトは，サイト・ライセンスや大学向けのディスカウントで対応している．しかし，サイト・ライセンスからカントリー・ライセンス（私立大学情報教育協会や文部省による一括契約）に拡大し，多くの大学で利用できるようにすることも模索すべきだ．（注：数値計算ソフト（NAGライブラリー）の例では，海外で州政府との一括契約という形態もでてきている．）

あるいは，LISP-STATのようなフリーソフトやStatPartnerのような廉価な商用ソフトの中から有望なものが育てば，機能と価格面で競合が起こり低価格化が促進されるだろう．

表4．統計ソフトの評価項目．

価格
稼働機種
GUIかCUIか？　特にグラフ
統計ソフトの使い分け
汎用統計ソフト
アルゴリズム記述
その他
表計算，数式処理
統計手法

(2) 稼働機種

稼働機種は，統計処理単独で考えるのか，統計を含んだアプリケーションとして考えるのかによって評価も分かれてくる．後者であれば，汎用機や WS の利用が考えられる．

しかし，統計教育やデータ解析は前者の代表であり，もはや PC で十分である．というよりも，汎用機で教育やデータ解析を行うことは，次の点で非効率であることははっきりしている．

- TSS や汎用機の OS に関するトラブルやそれらの知識習得の比重が高く，なかなか統計ソフトの利用に入っていけない．
- ファイルなどのマシン・リソースが制約され，利用者の閉じた世界で完結しない．すなわち，自由に利用できない．
- ワープロなどの他のソフトとのシームレスな連結が難しい．
- 会話形式のメリットが受けられない．
- GUI のメリットが受けられない．
- グラフ出力が弱い．あるいは，デフォルトになっていない．

国立大学における統計教育を指導している教官から，データ解析教育がうまくいかないとの声も聞くことがあるが，その主たる原因は汎用機の利用という制約によることが原因でないかと考えている．

(3) PC と Windows の問題

しかし，多くの大学教官の間では，PC と Windows による教育の管理上の難しさを指摘する向きもある．

- 故意あるいは，誤操作によるファイルや状態の変更に伴うトラブル．
- ウイルス感染や違法コピーの防止が難しい．

これらは，筆者も Windows を用いた 70 名を対象とする成蹊大学の情報リテラシー教育で一部経験しているが，管理方法の改善で解決できる．UNIX や汎用機での教育を肯定することにはならないと考える．今後 PC 利用による教育の運用方法については知恵のだしどころである．

(4) GUI か CUI か？ 特にグラフ

GUI か CUI か？ これは，データ解析教育を考えると，PC 上の GUI で行うのが次の理由でよい．

- 使用する際の敷居が低い．これによって，複数の統計ソフトを容易に使うことができる．
- グラフが容易に利用できる．これにより，統計量の意味を事前に直感的に理解でき，従来より統計をより親しみのあるものにできる可能性が出てきた．また，素人の誤用がさけられる．
- 他ソフト，特に結果をワープロと連動できる．

特にグラフは直感的に理解しやすく，統計ソフトを使った教育では従来の統計教育の内容をグラフを用いて視覚的に理解させる努力が必要であろう．

(5) 統計機能に対応した統計教育の多様性

統計教育を考えると，実データの解析技術とアルゴリズム教育は別物であることを認識すべきだ．

統計ソフトを用いた教育の形態としては，従来の数理統計による理論教育に加えて，

- 汎用統計ソフトによるデータ解析（主として，文科系学生）
- 行列言語によるアルゴリズム教育（主として，統計専門課程）

表5. 統計ソフトによる実験.

	SAS	SPSS	その他のソフト
学生の生活実態調査データ[1]	高森，新村（1987）	新村（1994b）	第9回日本計算機統計学会シンポジュウム
科学万博データ[2]	新村（1989）		
BANK.SAV[3]		新村（1995）	

[1] 40個体＊8変数，[2] 184個体＊19変数，[3] 474個体＊11変数.

・表計算ソフトによる簡単な計算練習（主として，文科系学生）
・数式処理ソフトを用いた統計シミュレーションによる統計あるいは確率分布理論の学習（主として，理工科系学生）

というような使い分けを模索すべきだ.

しかし，これらの特徴を生かし統計ソフトを使い分けるのはやはり大変であり，現在の商用ソフトを教育目的で使う立場からは欠点でもある.

(6) 統計手法

各統計ソフトの機能や手法の比較評価の情報が重要である.

しかし，新村（1995b）は，「学生の生活実態調査データ」という共通のデータを用いて共通の手法を用いた課題の発表を，SAS, SPSS, VisualStat, StatPartner, S, S-Plus, Juse-QCAS, Excel太閤，Mathematica, Mathcadの販売代理店やユーザーにお願いした．10頁までの予稿も作成してもらったが，それでも比較評価は難しい.

結局のところ，表5に示すように，幾つかの教育的なデータを用いてSAS（高森・新村(1987)，新村（1989)）とSPSS（新村（1994b)，新村（1995)）に関して，5.2で提案する解析方法で，比較評価を行ってきた．他の統計ソフトに関しても同様なことをさらに行う必要を感じている.

この試みで現在分かっていることは，

・用いる統計ソフトによって，大筋の方法論は変わらなくても，出力結果により説明の仕方や考え方が影響される.
・ここで考えた一般的な多次元データの解析手順以外に，医薬，マーケッティング，品質管理，時系列分析などの分野ごとの議論も必要となるだろう.

4. 第2世代で統計ソフトの果たした役割と問題点

第2世代における統計ソフトの果たした役割は，次の通りであろう．そして，多くの問題点が明らかになった．第3世代では，これらの問題を解決すべきであろう.

4.1 個人にとって

企業においては，日科技連や日本規格協会や統計数理研究所や業者セミナーなどによる統計知識の普及活動に社員を参加させているところもある．また，企業内で統計教育コースを独自に設定しているところもある．これらは恵まれた一部の企業である.

多くの企業で，このような講習会に参加できない者にとっては，私のように統計ソフトで独学し実践できたことが一番大きな貢献であろう.

すなわち，統計ソフトは社会人教育の役割を果たし，統計利用人口を企業で増やすのに貢献

した．

しかし，商用統計ソフトが導入されている産業や企業には明らかに偏りがあり，以上の効果はかなり個人的な努力と犠牲を伴う場合が多い．

4.2 企業にとって

産業応用としては，古くから医薬品業界や医学分野そしてマーケッティング分野等で，統計ソフトが使われてきた．

最近では，金融分野における投資分析システムや医薬の臨床比較試験システムのように，システムの一部として統計手法が使われている例も多い．

しかし，マーケッティングや医薬品業界等の統計処理が情報処理の主力になる一部産業を除いて，統計家は社内で評価されることが少ない．また，統計ソフトは情報処理部門が管理できない特殊なソフトとして扱われていることが多い．

すなわち，日本の企業の情報処理の主流は，まだ算盤の置き換えである単純な事務計算（集計計算）が多く，経営意思決定に統計・OR 等を実際に応用している企業は少ない．

これは，企業人や研究者に対する統計や OR などの効果をアピールできなかった我々にも責任はあろう．

4.3 大学における教育

一方，大学においては，研究者には役立っていても，情報処理教育や統計教育の道具としては，まだ十分に使われていないようだ．

例えば，筆者の関係している 2 つの大学では，恵まれたことに SAS が入っているが，あまり授業で積極的に利用されてきていなかったようだ．もう一つの大学では，表計算ソフトしか利用できない．

これから，次のような問題点が浮かび上がる．

- 今後の統計教育に適した統計ソフトとは何かを，もう一度考え直すこと．
- これまで費用の制約により，統計教育における統計ソフトの選択の幅を狭めてこなかったか？　もしそうであれば，どう改善するか．
- 統計ソフトを使った統計教育は，情報リテラシー教育との関係も含め，今後いかにあるべきか？

5. 如何に第 3 世代を築くべきか

以上で，第 2 世代の統計ソフトの功罪と残された問題点を述べてきた．これ以外に，次の点を検討し行動することによりこれからの第 3 世代を切り開くべきであろう．

- 第 3 世代の目標を明らかにする．
- データ解析の方法論を検討していく．
- データについての考えるべきテーマを明らかにする．
- 統計の理論と統計ソフトを用いた教育のありかたを考える．
- 固有技術との関係について考える．

以下では，現時点で筆者が考えていることを述べるが，データや固有技術に関しては十分検討していないので，本稿からは割愛する．

5.1 「素人の誤用」論を越えて「素人が間違いなく多用」できる時代へ

汎用統計パッケージの普及黎明期すなわち第2世代に入る初期で,「統計パッケージの素人による誤用」が指摘された.

これは,ある程度,警句として意味があったが,次のような点を検討し対応策を示せばもっと良かったと考える.

・素人が誤用しないための,対応策は?
・素人が誤用しないための,教育は?
・素人が誤用しないための,分かり易いテキストを含めたコースウエアは?

このような問題は,20年近くの統計ソフトの利用をふまえて,今日漸く議論すべき環境になったのかもしれない.

しかし,第2世代においても,SASのような成功した商用パッケージでは,ユーザー会が組織化され,事例発表や情報交換が行われた.ユーザー会は,学会の敷居の高さから疎遠になりがちな企業内の統計ユーザーの発表の場として有効であり,データ解析事例の啓蒙に役立った.

5.2 解析方法の標準化

筆者の持論であるが,「データ解析の手順の標準化」に関する議論が統計利用人口を増やし誤用をさけるためにも,重要である.

図3は,一つの提案である(新村(1994a)).このような簡単なものでも,初心者にとっては,

・解析目標と作業仮説の明確化
・データ収集方法の企画と検討(データの科学は,ここに焦点を当てるべきである)
・データの入力ミスの発見と新変数の誤りチェック

というような軽視しがちな重要問題を,習慣的に心がけるよう指導できる.

そして,表6(新村(1995))のように「統計手法を体系的にとらえ」,この順序である程度は機械的にデータ解析することができる.

統計手法の分類に関しては,林による「外的基準のありなしで始まる有名な体系化」がある.これは,数量化理論の位置づけの理解に便利であった.しかし,一般的な多次元データを素人がどう分析していけば良いかを示すものではなかった.

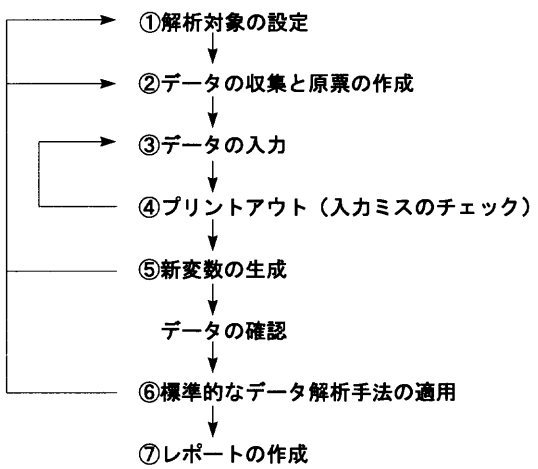


図3. データ解析の標準的な手順.

表6．統計手法のデータ解析としてみた体系化．

1)　データをとらえる

	量的変数	質的変数
1変数	ヒストグラム，幹葉図 基礎統計量	度数表
2変数	散布図行列と相関分析	クロス集計
3変数	主成分分析と主成分スコア クラスター分析	多重クロス集計

2)　予測手法

		目的変数	
		量的変数	質的変数
説明変数	量的変数	回帰分析	判別分析
	質的変数	箱ヒゲ図と分散分析	CHAID

表6は，単に手法を並べたようであるが，前にも述べたとおり，3つのデータを用いてSASとSPSSですでにこの主張が役に立つか否かを実験している．

しかし，医薬やマーケッティングなどの特殊な領域では，別の手順を検討する必要があろう．

5.3　グラフの影響

さらに，「グラフによる視覚的判断を統計量に優先させれば，初歩的な誤用はさけられる」と考えている．そして，新村（1995）で，米国の裁判所で争われた興味あるデータを用いて，これを実践してきた．図4は，その1例で，回帰分析のF検定をグラフ化したものである．このように統計理論の説明の前に，グラフによる直感的理解を優先すれば，難しいと敬遠されがちな統計も利用者が増え，誤用もある程度はさけられよう．

もう一度，従来の方法論や考え方や教育方法を再検討し，「素人の誤用」を越えて，「素人が間違いなく多用」できる時代へ我々が協力し先導すべきであろう．

5.4　統計ソフトを用いた統計教育

(1)　情報リテラシー教育との連携

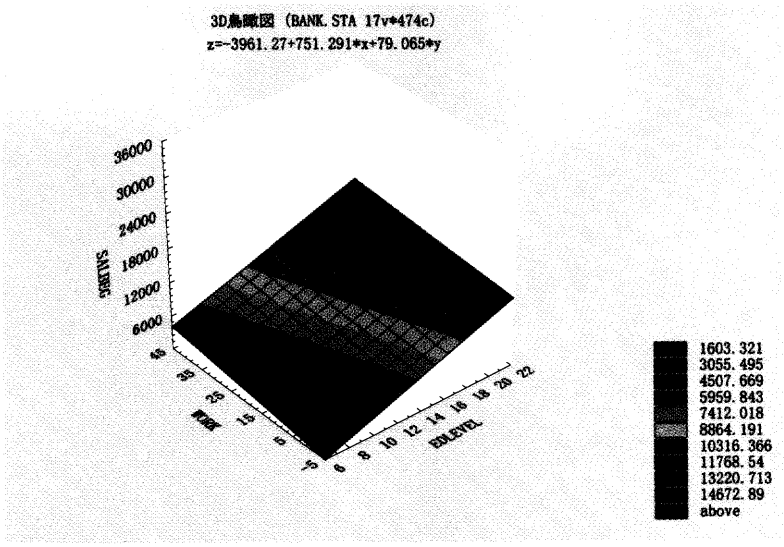
筆者は，一つの提案として「統計やOR教育の一部を，情報処理教育とドッキングしたデータ解析教育（あるいはデータの科学）やモデルの科学の教育として行うべき」と考えている．

統計は，従来は記述統計，推測統計に分類されてきたが，ここにきて統計ソフトの支援を受けたデータ解析が第2世代でつけ加えられた．

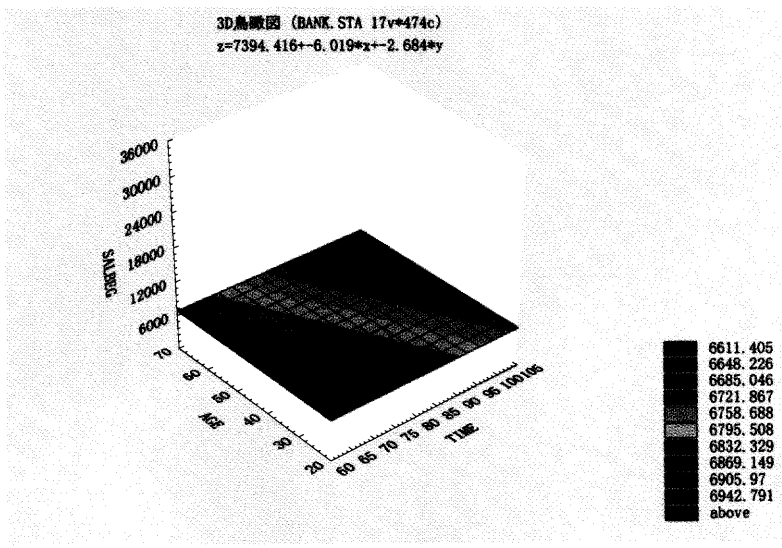
情報処理教育の内容を考える上で，パソコンなどの紹介本で言われているのは，次のようなものである（石田（1988））．

・ワープロ
・表計算
・データベース
・通信とインターネット

これに加えて，大学ではコンピュータ言語教育が行われている．果たして情報リテラシー教



(初任給を学歴と就業年数で予測：意味のある回帰分析)



(初任給を年齢と熟練度で予測：意味のない回帰分析)

図4． 回帰分析のF検定の図的理解．

育はこれでよいのだろうか．

少なくとも，データとモデルをどう扱うかを考える学問として，データ解析（あるいはデータの科学）とOR（モデルの科学）を情報処理教育に取り入れる必要があるのではなかろうか．

また，対象を文科系と理科系，情報処理の専門家と利用者という立場から，もっときめ細かに分けて，その教育方法を検討する必要もあろう．

このため，今後は，教育的なデータについての研究も必要となる．

(2) データ解析教育の問題点

すでにデータ解析教育を行っている場合でもうまくいかないという声が多いが，はっきりし

表7. アンケート調査の項目.

使用機種
使用ソフト
使用データ
対象人数と年次(早い方がよいと考える)
前提の知識と他の統計教育との関係
教育法・教材・カリキュラム
教える側の背景

た情報がないため理由が分からない．これにたいしては，現在行われているシラバスなどの調査の必要があろう．また，表7のようなアンケート調査や，学会でセッションを設け議論する必要があろう．

(3) 教材としてのデータ

一方，データ解析教育に用いるデータが無いという声もある．これは，理由にならない．統計ソフトには種々のデータが含まれている．例えば，新村(1995)はSPSSに含まれるBANK.SAVというサンプル・データを用いた解析事例である．しかし，せっかくのサンプル・データも詳しい説明が無い物が多く，提供者側の認識の甘さも目に付く点である．

また，筆者の非常勤講師先の東洋大学の1年次の情報処理概論で「自分で100件以上のデータを集め，SASを使って解析し，結果を一太郎でレポートしなさい」という自由課題で，400人中100人ぐらいは種々の経済データを図書館で探して，評価できるものを提出した．すなわち，データ解析の対象となっていない生データも多いのが現状である．

しかし，今後はいろいろな分野における特徴あるデータを広く収集し，教材としての適否を検討すべきであろう．

また，有用なデータと解析事例を整備し社会的に還元していく必要がある．

5.5 ネットワーク技術との共生

今後，第2世代に決別し第3世代を築いていくためには，ネットワーク技術と共生する事が考えられる．

(1) 情報発信とWWW

これまで，企業人は，統計に関する情報を書籍やセミナーや統計ソフトのユーザー会を通じて得てきた．研究者は，さらに学会活動などで，最新の情報を得てきた．

そこにおける問題は，情報を発信する側も得る側も，費用や時間の制約やアクセスに問題があった．これをある程度解決してくれそうなのが，WWWであろう（今泉（1996））．

今後は，

- 企業や組織での利用から，個人利用の促進
- 統計ソフトを使った教育法の研究と普及
- 教育用のデータの整備と公開
- 統計知識，利用法を調査・整備し，広く公開し普及に役立てる事

が必要になる．

このための方便として，次のようなことを行うことを提案したい．

- シラバスを含め現在の統計関連と情報リテラシー教育の幅広いアンケート調査の実施

・有用なデータの収集と登録を行い，公開する
・統計知識の共有化を計る
・ユーザー事例の登録制度を検討する
・教育・セミナーの調査を行う（新家 他（1996））
・商用ソフトの開発企業や国内外の大学研究機関のホームページの調査

そして，これらの成果をインターネットの利用を通して普及し発展させていくことを検討すべきであろう．すなわち，誰もがアクセスできる統計ホームページを公的に作成し，次のような内容の情報公開を考えてみてはと思う．

・国内外の統計関連学会のホームページとのリンク
・各統計ソフト・ベンダーのホームページとのリンク
・国内外の主要研究室や機関のホームページとのリンク
・利用者教育の実例の登録
・ソフトやデータや解析事例の調査結果の登録
・共有データの公開
・ノウハウの公開
・統計相談コーナーの開設

ノウハウに関しては，最近ではWWW上で稼働するAIシェルもあるので，多くの研究者の知恵を統合化し，インターネットとAIによる知識の共有化を計ることを模索すべきであろう．

これを行う実施主体は，統計関連学会，統計数理研究所，研究者の個人研究と，どこにするかは検討の余地がある．しかし，すでにある統計関連学会ホームページ(http://sunyht2.ism.ac.jp)を充実させるのが本筋であろう．そして，これらの成果をWWWサーバー上に公開し，広く国民により広い知識の提供と共有を実現すべきである．

(2) AI再考

統計の分野でも，AIの研究が注目を集めたが，華々しい成果は上がっていないようだ．また，英国NAG社によるGLIMの利用法を，Prologでコンサルテーションする商用のエキスパートシステムのGLIMPSEのサポート中止の例もある．

AI化がうまくいかない理由としては，次のことが考えられる．

・AIシェルから作り始める．あるいは，LISPやProlog等の言語を用いて一から開発する．しかし，これでは開発に時間と費用がかかりすぎる．プロトタイプの作成は，AIシェルで開発すべきであろう．
・統計エキスパートシステムの多くの開発者の興味は，統計やデータ解析より，情報処理システムに主体があったのではなかろうか．これまでの多くの開発者に，本当にデータ解析の実務経験があったのだろうか疑問である．
・選んだテーマが適切だったか疑問である．
・知識の保守と更新は，少人数ではできにくい．RSA暗号の解読の例として，インターネットで作業分担した成功例もある．最近では，WWW上で動くAIシェル(例えば，EXSYSのhttp://www.exsysinfo.com参照)もあるので，各研究者が作業分担し共同で統計エキスパート・システムを作成することが良いように思う．

6. 終わりに

これまで，「数学は，自然科学の共通言語」として位置づけられてきた．これに対して，我々

の分析したい対象は，データやモデルとして記述される．データを科学的に扱う有力な手段が統計であり，モデルを扱うのがORである．これらの学問を，従来の理論偏重の単独の学問分野として考えるのではなく，ソフトを利用することにより，データの科学やモデルの科学として再編し，多くの学問を共通にとらえ総合化するための基礎学問（触覚）にしなければいけない．

そして，大学教育を，技術者や研究者向けの理論中心の講義に加えて，情報技術を利用した情報リテラシー教育を充実し，卒業生が産業界で応用できるよう再検討すべきであろう．

参考文献

芳賀芳郎（1996）．『データ解析の基礎』（日科技連MA基礎コーステキスト），(財)日本科学技術連盟，東京．

Hayashi, C. (1996). What is data science ?—Fundamental concept and heuristic examples—, *IFCS-96*, Abstracts Vol. 1, 53-56.

今泉　忠（1996）．統計学におけるインターネットの利用，第64回日本統計学会講演報告集，60-61．

石田晴久（1988）．『パソコン入門』，岩波書店，東京．

大森　崇（1996）．動物実験代替法バリデーション研究のデータマネジメント，第64回日本統計学会講演報告集，54-55．

大隅　昇，秋山直樹（1988）．問題だらけの国産統計ソフトウエア，日経バイト，**AUGUST**，225-239．

関田素子（1995）．VisualStat (ver5.0j) による統計解析の視覚的アプローチ，第9回日本計算機統計学会シンポジュウム論文集，35-44．

柴田里程（1996）．データ・サイエンスにおけるデータベースの役割，第64回日本統計学会講演報告集，68-69．

新家健精，竹内　潔，松下嘉米男（1996）．産業界における統計教育について，第64回日本統計学会講演報告集，180-181．

新村秀一（1989）．『やさしく実践データ解析の進め方』，共立出版，東京．

新村秀一（1993）．『意思決定支援システムの鍵』，講談社，東京．

新村秀一（1994a）．『SAS言語入門』，丸善，東京．

新村秀一（1994b）．『SPSS for Windows入門』，丸善，東京．

新村秀一（1994c）．Mathematicaによるデータ解析へのはじめの一歩—統計パッケージとの比較検討—，Mathematicaユーザー会第2回ワークショップTokyo 1994講演予稿集，80-89．

新村秀一（1995a）．『パソコンによるデータ解析』，講談社，東京．

新村秀一（1995b）．第9回日本計算機統計学会シンポジウム開催に際して，第9回日本計算機統計学会シンポジウム論文集，7-16．

新村秀一（1996a）．統計ソフトと情報リテラシー，第64回日本統計学会講演報告集，58-59．

新村秀一（1996b）．重回帰分析と判別分析のモデル決定（2）—19変数を持つC.P.D.データのモデル決定—，成蹊大学経済学部論集，**27**(1)，180-203．

新村秀一，三宅章彦（1983）．重回帰分析と判別分析のモデル決定（1）—19変数を持つC.P.D.データの多重共線性の解消—，医療情報学，**3**(3)，107-124．

高森　寛，新村秀一（1987）．『統計処理エッセンシャル』，丸善，東京．

垂水共之，山本義郎（1996）．データ辞書を用いたデータ分析システム，第64回日本統計学会講演報告集，66-67．

椿　広計（1996）．データの品質とそのマネジメント，第64回日本統計学会講演報告集，52-53．

上田尚一（1996）．統計教育用ソフトの問題点，第64回日本統計学会講演報告集，189-190．

矢島敬二，大隅　昇（1977）．統計—統計ソフトウエア—，「ビット」別冊—アプリケーションプログラム—，**9**(9)，894-920．

吉田典弘，石橋雄一（1995）．新統計教育用のソフトウエア（2）—StatPartner多変量解析オプションを活用した学習—，第9回日本計算機統計学会シンポジュウム論文集，63-72．

A New Step from Data Analysis to Data Science —How can We Spread Statistical Knowledge to Our Nation in Collaboration with Statistical Software, WWW and AI Technologies?—

Shuichi Shinmura

(Department of Economics, Seikei University)

I think there are three generations of statistical software.

First generation is the age of bespoken software developed by individual researcher. It is the age of craftsman. Its aim is to help research of statistical theory.

Second generation is the age of commercial statistical package such as SAS and SPSS. These packages open new world of "Data Analysis" of actual data in many fields such as medical care, medicine, marketing, engineering, management and research.

Now, we are at the entrance of the third generation. What is our target and what shall we do? We had better set a high target that we do our best for most people to use statistical knowledge in usual work and research in addition to word processing and spreadsheet.

In order to attain this target, we develop new frontier named "Data Science". "Data Science" needs a good balance of statistical package usage technique and know-how of data analysis and deep knowledge of data itself.

People say mathematics is a common language of science and technologies. I think "Data Science" should be a sensor of many fields in addition to these area. In order to open new world, we must change our old style of statistical education in universities and spread new wisdom of "Data Science" by combinations of WWW and AI shell such as EXSYS.

Key words: SAS, SPSS, three generations of statistical package, computer literacy, Data Analysis, Data Science, WWW, AI.