統計パッケージ考——情報処理教育と産業——

新村　秀一

本稿では、統計ソフトの中で特に汎用統計パッケージといわれるものを取り上げる。そして、統計教育や産業的な視点から、その影響を考える。

1　パッケージとは

パッケージとは、なんであろうか。そして、その果たす役割と社会への影響は、どうであろうか。

（1）コンピュータ言語

この問いに答えるため、コンピュータのプログラミング言語の歴史を振り返ってみよう。

ご存じのように、コンピュータの言語は、第1世代の機械語、第2世代のアセンブラー、第3世代の高水準プログラミング言語（The 3rd generation language、3GLと略す）というように発達してきた。

これは、コンピュータ・ハードウエアの処理能力の向上に対応して、より高機能なものが開発されてきたからである。コンピュータの負荷（冗長さ）を犠牲にして、プログラマーに使いやすさをもたらした。これによって、著しくプログラミングの生産性が向上した。

（2）サブルーチンとライブラリーによる生産性向上

次ぎに、サブルーチンとかライブラリーと呼ばれるよく使われるプログラムを部品化し、再利用することによって、生産性を向上させる方策が考えられた。

このような部品化は、技術計算の分野では成功したと言えよう。その証拠が、商用のFortranライブラリーである。IBMから独立して作られた米国のIMSLや、英国の非営利研究開発企業NAGのライブラリーなどである。

これに対して、COBOLを用いた事務処理の分野では、ワーニエ法やオランダのダイクストラが提唱した構造化技法などが流行し廃れていった。

科学を土台にした技術計算では、標準化や部品化が容易であり意味があるのに対し、企業における個別の約束事を標準化や部品化するのは始めから難しい点がある。この点は、あまり指摘されていないようだ。

最近はやりのオブジェクト指向は、COBOLに代わって、新しい言語C++などによって、クラス・ライブラリーすなわち部品を作ろうという試みである。

以上述べた生産性の向上の試みは、プログラマーと呼ばれるそれを職業とする専門家のための生産性向上のための歴史である。

（3）パッケージ

これに対して、今注目のダウンサイジングによって、生産性向上の試みはユーザーをも巻き込んで別の展開をしてきている。

ソフトウエアは、プログラミング言語とそれによって開発されたOS、アプリケーション、ユーティリティに大別される。

OSは、ハードウエアとアプリケーションの間にあって、これらを有効に運用管理するシステム・プログラムである。ユーティリティは、ちょっとした共通に使われる便利なプログラムである。後で述べるパッケージ的な側面をもっている。アプリケーションは、特定業務用のソフトウエアである。はじめは、FortranやCOBOL等の3GLで開発され、その多くは一品生産のオーダー・メイドのことが多かった。ユーザーに密着したソフトウエアである。英国ではbespoken softwareという。

しかし、その中から汎用アプリケーションとかパッケージ・ソフトウエア略してパッケージと呼ばれるものが作られるようになった。パッケージとは、包みをとくだけで誰もが簡単に利用できる事を意味している。

すなわち、パッケージは不特定多数のユーザーを対象とし、簡単なコマンドで操作できるものをいう。

蛇足であるが、このようなパッケージを作るソフト

しんむら　しゅういち　住商情報システム（株）
〒130　墨田区両国2丁目10番14号　両国シティコア
☎03(5624)1731　FAX03(5624)1725

ウエア会社と、大規模なオーダーメイドのソフトウエアを開発する企業とは、別種のソフト産業である。そして、アメリカにおいてはパッケージ産業が優位にあり、日本では後者が優勢である。

このようなパッケージの歴史において、統計パッケージは、他の分野に比べて先行した歴史をもっているし、一応成功していると言えよう。

2　統計パッケージの効用

（1）統計パッケージ利用のメリット

統計パッケージ利用のメリットは何であろうか。それは、自分で統計プログラムを作成することを考えてみればよい。

少なくとも統計アルゴリズムとプログラミング言語とコンピュータのシステム的なことを、かなりのレベルで理解していなければならない。そして、デバッグを行った後、ようやく自分の仕事に利用できる。

しかも、統計手法といっても色々ある。とても自分で一から作っていたのでは、生産的でない。人は仕事量にあわせて、自分の仕事を制限しがちだ。私が社会人になりたての頃、コンピュータの費用が高く、統計パッケージ利用が一般的でなかった。そこで因子分析や重回帰分析を一回実行すれば、それで一仕事あるいは論文が一丁完成したわけである。

これ以外に、ソフトウエアの信頼性という問題がある。自分で自分を信用できないのは情けないことだが、間違いのないソフトウエアを作る労力は大変なものである。また、開発よりも保守が重要だ。

（2）プログラミング言語とパッケージ

表1は、プログラミング言語とパッケージ（エンドユーザ言語）を別の観点からまとめたものである。

表1　プログラミング言語とパッケージ

世代	言語	水準
第1世代	機械語	電気信号のレベル
第2世代	アセンブラ	同上
第3世代	Fortran,COBOL,BASIC	プログラムのレベル
第4世代	パッケージ	仕事のレベル

すなわち、機械語とアセンブラは、0/1の電気信号のレベルでコンピュータに命令する言語である。

これに対して、3GLとして最も歴史の古いFortranや、ダートマス大学の数学者ケメニーらによって作られたお馴染みのBASIC等は、人間の思考にあったプログラムのレベルといえよう。ここまでは、ある程度以上の適性を備えたプログラマによって用いられる。

これに対して、パッケージは、ある仕事をコンピュータによって行いたい全ての人のための、いわゆるエンドユーザ言語である。

3　統計パッケージの歴史

次ぎに統計パッケージの歴史を見てみよう。

3．1　米国の動向（汎用機御三家）

日本人は、戦後全てにわたってアメリカに顔が向いているせいか、統計パッケージといえば米国の汎用コンピュータ用のBMDP(Bio Medical Computer Program)、SPSS(Statistical Package for the Social Sciences)、SAS(Statistical Analysis System)が日本では有名であった。

（1）BMD

1956年にカリフォルニア大学の医学部で、BMDと呼ばれる商用の統計ライブラリーが開発された。医学は、統計が重要な分野であり、共通の財産として統計ライブラリーが開発された事はよく理解できる。BMDは、統計ライブラリーであったらしいが、その後BMDPと呼ばれる統計プログラムに集大成されたようだ。このあたりは、実際に利用した経験が無く、文献などのまた聞きである。日本では、東京大学大型計算機センターなどで使われていたようだ。代理店がないので、直接アメリカから導入したのであろう。

（2）SPSS

SPSSは、1965年にスタンフォード大学の社会学部で開発された。現在では開発者が設立したSPSS Inc.（シカゴ）が開発サポートしている。名前が示すとおり、社会科学向けに作られたが、現在では分野にとらわれない、汎用統計パッケージである。

日本では、当時京都大学大型計算機センターの山本先生や、北海道大学の司馬先生らが、代理店がないにもかかわらず啓蒙活動の一貫として、解説書を出版されたことは特筆に値する。

このころに、一部の統計学者の間で、「素人が便利な汎用統計パッケージを使うことによる誤用の危険性」という有名な議論がなされたようだ。

（3）SAS

SASは、1966年に、ノースカロライナ州立大学の統計

学者A.J.Barr氏やJ.H.Goodnight氏（現SAS社社長）らによって、開発された。

Goodnight氏は、統計学科を卒業後、GEに入社し政府プロジェクトに参加し、制御関係の開発に携わった。その後ノースカロライナ州立大学の統計学部に戻り、上記の経験を生かして大学でSASを開発して、1972年にSAS72版を大学にリリースした。1976年に大学から独立し、SAS Institute Inc.を設立し今日に至っている。

3．2　ヨーロッパの動向

華々しいアメリカの動向に対して、ヨーロッパの統計パッケージや動向については、文部省統計数理研究所の大隅氏らがヨーロッパ、特に英国のNAGとフランスのパリ大学における数量化のソフトウエアを紹介をしているのが先鞭である。

Genstat5は、実験計画で有名なフィッシャーがいたロザムステッド実験農場で開発された汎用統計パッケージである。GLIM(Generalized Linear Interactive Modelling)は、英国王立統計協会で作られた、有名な一般線形モデルの専用統計パッケージである。これらは、その後NAGが開発・販売している。

3．3　日本の動向

日本においては、各汎用機メーカーが汎用統計パッケージを開発していたようである。

昭和46年に現在の会社に入社した筆者は、医療データ解析のためにNECのSTAT-EXと呼ばれる汎用統計パッケージのお世話になった。他のメーカも、同様なものを開発していたようだ。

一方、大学では、当時九大の浅野先生らを中心に文部省の科研費により開発されたNISANシステムがある。また、文部省統数研では、赤池先生らの業績を具現化した時系列パッケージのTIMSACがある。これらは、国の予算を使って開発されたので、商用化に制限があった。アメリカにおいては、多くのソフトウエアが大学で開発され、商用化されたのと比べて際だっている。

このほか、日科技連では、OR、品質管理、統計の企業への普及を指導してきており、この関係で商用の統計パッケージを開発し販売している。

3．4　新しい流れ

統計ソフトは、今大きな変革期にある。それは、パソコンの処理能力が上がってきたために、統計処理（データ解析）は、パソコン（ＰＣ）でも十分な時代になってきたからである。

PC用の統計パッケージの利点は、次ぎの通りである。

・PCの処理能力がひと頃のWS並になった。そうであれば、WSよりもPCのほうが、流通ソフトの豊富さ、Windowsの使いやすさ、価格の安さなどから優位になる。

・機能とは無関係に、PC用のソフトの価格は安くせざるをえない。このため、個人でも入手できる。

以上から、今後はPC市場での汎用統計パッケージが主流になっていくだろう。

現在この分野では、SASやSPSSやGenstat5などの汎用機から降りてきたものと、PC用に新規に開発されたものとが混戦状態にある。

SASは、そのままの設計思想、販売政策をPCに持ち込んでいる。従来のユーザであれば、熟知しているSAS言語をより使いやすいPCの環境で使える。価格は安くなったとはいえ、レンタル性は踏襲されている。このため、新規ユーザが、導入することは少ないだろう。

これに対して、SPSS for Windowsは、従来のSPSSのコマンドの上に、Windowsのアイコンを重ねてオブジェクト指向に脱皮している。また、PC用のソフトに一般的な売りきりである。

一方、PC専用に開発された汎用統計パッケージとしてVisualStat、Jump、SigmaStatなどがある。

以上述べた汎用統計パッケージと異なる動きとして、AT&Tで開発されたＳがある。統計の頭文字を採って、Ｓと付けたのであろう。これは、関数型のプログラム言語である。スカラー処理だけでなく配列や行列を扱えるので、アルゴリズムの記述に適している。Ｃと同じく、安いロイヤルティで公開されているので派生バージョンが幾つかある。S-PLUSは、Ｓに独自の関数を追加している。

4　統計手法とソフトの分類

4．1　統計ソフトで何をするか

筆者は、統計手法を大きく分けて、表２の「データの分布を調べる手法」と、表３の「予測手法」の２つに分類して考えることにしている。そして、それらの統計手法と関連したグラフ手法が必要になる。

汎用統計パッケージという場合、最低でもこれだけの手法を提供すべきである。

（１）分布を調べる手法

分布を調べる手法では、数値変数かカテゴリー変数かの軸と、変数の数の２つの軸によって、６つのカテゴリーに分かれる。

１個の数値変数では、まずヒストグラムや幹葉図や箱ヒゲ図で、分布の特徴をつかむ必要がある。そして、

次ぎに正規性の検討を行い、基礎統計量や各種の検定統計量の意味を考えることになる。

２個の数値変数では、散布図で２変数の関係を概観し、相関や偏相関を検討する。

３変数以上の数値変数は、主成分分析、因子分析やクラスター分析で検討される。

一方、カテゴリー変数では、単純集計や多重クロス集計が必要になってくる。

表２　分布を調べる手法

	数値変数	カテゴリー変数
１変数	正規性の検定 基礎統計量	単純頻度
２変数	相関と散布図	２重クロス集計
３変数 以上	主成分分析 因子分析	多重クロス集計

（２）予測手法

予測手法は、目的変数と説明変数の２つの軸で、数値変数かカテゴリー変数かの違いを考えることにより、４つのカテゴリーに分かれる。

重回帰分析が特に重要だ。

表３　予測のための手法

		目的変数	
		数値変数	カテゴリー変数
説明変数	数値変数	重回帰分析	判別分析
	カテゴリー変数	共分散分析 分散分析	多重分割表

（３）その他

以上が、一般的な汎用統計パッケージが備えていなければいけない手法である。このほか、時系列解析や品質管理や検定などの手法をオプションでもっていることが望ましい。

４．２　統計ソフトのスペクトラム

（１）ライブラリー

IMSLやNAGのようなライブラリーには、統計手法が含まれている。ライブラリーである以上、あれもこれもと出力することは間違っている。このため、必要最低限の出力に限られている。回帰分析を例に取れば、回帰係数とか分散分析表程度がライブラリーの出力である。詳しいモデル診断は出力されない。

（２）中間言語

中間言語という言い方は、どこかで聞いた記憶があるが覚えていないので、ここでは私の造語としておく。行列や配列を扱うことができるので、行列言語という人もいる。スカラーを扱う3GLとエンドユーザ言語の中間に位置するプログラミング言語である。

例えば、AT&Tで開発されたＳ言語がある。パッケージで提供していない手法がある場合、Fortran等で一から作ることは大変だ。もともと、統計手法の多くは、行列で記述できるので、中間言語を用いればプログラミングが容易である。出力は、やはりライブラリーと汎用統計パッケージの中間程度である。

（３）汎用統計パッケージ

汎用統計パッケージの特徴は、４．１で述べた手法をサポートし、しかもかなり詳細な出力が得られる点である。このほか、

・外部データの入出力と編集加工
・標準ファイルの管理と操作
・連続処理
・文法チェック

等が最低必要である。そして、最低でも数千件程度のデータ処理が行えるべきであろう。

一般的な情報処理の基本は、データの検索と更新やファイルの連結である。このためのツールとして、DBMSがある。表計算ソフトでは、情報処理の重要なこの機能を教えることが難しい。それは、変数とレコードという概念がないためであろう。汎用統計パッケージは、DBMSに比べて十分でないが、変数とレコードという概念があり、これを教えることができる。

（４）4GL

SASは、SPSSのような汎用統計パッケージの機能の他、開発言語の機能を持っている。いわゆる第４世代言語である。元々商品コンセプトが、「All in one system」すなわちSASひとつで全てをカバーしようという戦略である。初期の頃は、私自身大いにこれを喧伝したが、ダウンサイジングの時代にあって、以下の問題がある。

・統計ユーザにとっては、開発言語の機能はいらない。
・使わない分までの使用料を必要とする。

SASの問題は、使用料の高さであろう。ただし、大学でのサイトライセンスは意外と安いようだ。

（５）周辺ソフトによる統計処理への疑問
（膨張主義を排す）

最近の統計分野におけるトピックスとして、ハーバード大学がSASをキャンセルしたことである。理由は、統計教育も表計算ソフトで十分ということらしい。日本においても統計研究家の中で、統計の専門家以外には表計算ソフトで十分という人もいる。

筆者は、これには異論がある。将来はともかく、今の表計算のレベルで、満足な統計の理解がえられるのであろうか。今後おおいに議論すべきである。

たしかに、表計算は利用人口が多く、安くて便利である。イントロとして、統計や他のこともできるという程度の紹介は必要であろう。しかし、纏足のように伸びる能力を不自然に矯正する事にならなければと杞憂している。しかし、最近の汎用統計パッケージは表計算ソフトのデータを入力できるので、この連動は便利である。

このような一つのソフトで多くの分野をカバーしようとする膨張主義は、表計算に限らず売れ筋のSASやMathematica等に見受けられる。利用者の冷静な判断が必要だろう。

（６）良いソフトウエアとは

良いソフトウエアとは、教科書レベルの玩具の例題が解けることではなく、企業レベルの実用的なものが少ない労力で解けることである。このような視点さえもっておれば、間違った選択は避けられるだろう。

５　なぜ日本やメーカーは駄目なのか

汎用統計パッケージに限らず、DBMS、通信・ネットワーク、表計算などの多くのパッケージがアメリカ製である。なぜ日本から育たないのだろうか。

（１）なぜ日本が駄目なのか

答えははっきりしている。教育の問題もあるが、これらは始めから、世界市場を対象にして開発し、販売されるべきものである。この単純な認識が欠けていることに問題がある。

最近のアメリカからは、湯気の立つソフトが日本に代理店を求めてやってくる。しかし、始めから世界を相手にという意気込みのパッケージが日本にあっただろうか。一太郎が健闘しているのは、日本語ワープロという特異環境だからである。

（２）なぜメーカーが駄目なのか

なぜメーカーが駄目なのかは、メーカーの狭い市場ではもう駄目だということのほか、次ぎの問題点があるようだ。すなわち、ハードウエアが主であり、ソフトウエアが従である限り、ソフトがハードのおまけである限り、サード・パーティのソフトに勝てないということである。

メーカではハードの改良には心血が注がれるが、一旦パッケージができるとハードほど改良に次ぐ改良ということにならず、組織が縮小されるようだ。

６　情報処理教育と情報処理産業

筆者自身、初等・中等・高等教育に意見を述べる見識はない。しかし、大学教育は、一部の研究者を除いて、社会人になる前段階であるから、多くの社会人が意見を言い、議論すべきだと思う。

大学教育は、少なくとも研究者教育と実務教育を明確に区別すべきであろう。

そして、情報処理教育をつまらないカリキュラムでお茶をにごしてはいけないと思う。

情報系を除く理工学部や文科系の学部の情報処理教育は、統計を３割、できればORを２割、数式処理や可視化技法を各１割程度、一流の汎用パッケージを用いて教えるべきであろう。

骨董品の鑑定の世界で言われていることであるが、決して偽物を見てはいけないと言うことである。目の肥えた客を育てないと、日本からいつまでたっても一流のパッケージ・ソフトは誕生しないように思う。

このことは、江戸時代の商家の婦人が、明治の生糸産業の礎になったことでも歴史的に証明されている。

情報処理産業を21世紀のリーディング産業に育てようと言う試みは、今までのところことごとく失敗に終わっている現実を直視する必要がある。

また、専門家向けの情報処理教育もしっかり考えないと、筆者のような外国のパッケージの評論家が、21世紀になっても同じ事を言っている姿を想像しただけで、目の前が真っ暗になる。

＜参考文献＞

[１]新村秀一(1993):「意思決定支援システムの鍵」、講談社
[２]新村秀一(1994):「SAS言語入門」、丸善
[３]新村秀一(1989):「易しく実践データ解析の進め方」、共立出版
[４]新村秀一(1995):「SPSS for Windows入門」、丸善
[５]真鍋龍太郎、逆瀬川浩孝、若山邦紘(1988):「文化系のコンピュータ／応用編ー表計算ソフトの活用ー」、岩波書店