

StatPartner 開発の背景とその活用

NEC ソフトウェア* 吉田 典弘・工藤 重人

(1996年6月 受付)

1. はじめに

インターネットやパソコンなどの情報処理環境の急速な普及に伴い、多変量解析をはじめとする統計手法は幅広い層で利用される気運が高まって来ている。本文においてその中で特に統計手法を使い始めようとする層に焦点を当て、その背景と状況を概観する。そして、そのような層に焦点を当てた統計解析ソフトウェア StatPartner について紹介する。

2. 統計手法の利用現場を取巻く状況

本章では、ビジネス分野、教育分野、研究分野に焦点をあて、統計手法を使い始めようとする層の状況と課題について概観する。

2.1 ビジネス分野

規制緩和や価格破壊が進む中、マーケット構造や消費者ニーズなどの予測は近年ますます困難な状況になっている。これは誰も経験した事のない変化が次々に起り、経験に裏付けられた勘が生かされない時代になっているためである。従って、科学的な根拠に基づいた意思決定の手段が求められている。

このような状況の中、イントラネットやデータウェアハウスのようなキーワードが飛び交い、いわゆる意思決定支援システムに類するシステムの構築が脚光を浴びている。意思決定支援システムは古くから提唱され試行されてきた。そして名称も MIS (Management Information System) から始まって DSS (Decision Support System), ESS (Executive Support System) などのように変化してきたがその本質は蓄積された情報を IT (Information Technology) を利用して企業の意思決定に活用しようというものである。そして近年の動向としては、情報系データベースとその分析系システムをイントラネットや多次元データベースなどの最新の IT を元に構築しようというものである。

統計手法では「予測」「分類」「分析」「判定」などを目的としており、従来の経験が生かせない現在、多くのビジネスマンにとって統計手法のこれらの機能は非常に魅力的に見える。従って、意思決定支援システムの分析系に占める統計手法の役割は大きなものになっている。

一方、多くの企業においては統計手法をどこまで適用するか試行錯誤しながら進んでいるのが現状である。このような中、統計手法を有効活用できる環境の整備とシステムを構築できる人材の育成、システムを利用する側のレベルアップが急務になっている。

2.2 教育分野

統計手法は大学や企業などで教えられているが、ここでは統計手法を教える「統計教育のカ

* プロフェッショナル・ソリューション事業部：〒136 東京都江東区新木場 1-18-6.

リキュラム」と統計手法の教育を含んだ「情報リテラシ教育」について述べる。

(1) 統計教育のカリキュラム

統計手法を教える場合、分散や分布についての考え方からはじまって多変量解析などの理論までを教える。そして最後に解析実習を行い理解を深めさせる、というのが一般的な流れであろう。しかし、実際は計算手順の学習に終わっているケースが多く見受けられる。これは、カリキュラムに充てる時間が限られている事や、実習を行う環境が整っていないことに起因すると考えられる。いずれにせよ、データの集め方や結果の見方までを教えないと、実際に使える知識にならない。従って、カリキュラムを効果的に進める事のできる環境の整備が必要である。

(2) 情報リテラシ教育

以下に述べる教育を示す言葉としてはコンピュータリテラシ教育と言った方が適切かもしれないが、ここでは情報処理の環境を使いこなす課題を解決できる人材の育成と言う意味で話を進める。情報リテラシ教育について従来は、「パソコンの操作をどこまで教えるか」といったレベルの議論もあった。しかし、パソコンやインターネットなどの情報処理の環境が急速に普及するなか、情報処理の環境と専門知識を使って、「いかに情報を収集し」「収集した情報をどのように分析するか」までを教える必要性が高まっている。

たとえば、マーケティングのデータがあった時、従来その分析は専門家に依頼していた。これは、分析環境が手に入りにくい事や、数ヶ月かけて分析しても問題がなかったからである。しかし現在では、分析環境は個人でも手に入れることができるし、分析結果を直ぐに入手し次の意思決定に生かす必要がある。つまり分析結果を必要とする人が自ら分析する必要性と可能性が高くなったわけである。

また、知識を入手するためにインターネットで接続されている Web のサイトを渡り歩くと非常に多くの知識を入手する事ができる。そして新たな知識は次の知識を要求し、知識の広がりにはどんどん大きくなる。従来は、新たな知識を得るためには文献の調査が主たるものでありその意味で限界があった。しかし今ではインターネットを通じ容易に新たな知識にたどり着くことができるのである。これは従来、分野と分野の間にあった壁が取払われることを意味する。そして、たとえばマーケティングの専門家でない人もマーケティングの分野に足を踏み入れ、アンケートデータを分析することもできる。

このような状況に合わせた統計手法に係る教育は、「統計手法の基礎を教える」というものではなく「データから情報を抽出するための方法」を教えることが中心になる。つまり統計の基礎から学習するのではなく、統計手法の基礎はブラックボックスとして統計解析ソフトウェアの中に封じ込め、その入力と出力を押え、統計手法の機能と結果の利用方法を中心に理解させる、というものになると考えられる。もちろん「データから情報を抽出するための方法」は統計手法だけでない。データの可視化からはじまって、最適化・シミュレーションなどの方法も、今日の情報化社会を生きるための手段として教えるべきである。なおこれらの教育は文系・理系を問わず、身につけるべき情報処理のテクニックの一つである。

しかし、現状では教育環境とこれらを教えられる人材、カリキュラムの整備がなされていない事により、まだ十分な教育がなされていないのが実状である。

2.3 研究分野

統計手法はデータを扱った研究では、研究の当りをつける局面からはじまり、研究を深める局面、研究成果を裏付ける局面などに有効に利用できる。しかし、必ずしも全ての研究者が有効に統計手法を活用しているとはいえない状況である。実際、論文発表直前になって、「論文の結論が正しいか統計処理して検証してほしい」といった形での依頼を時々受ける。各研究者と

も統計処理の重要さは認識しているが、その環境がなくこのような事態に陥っているようである。統計処理を行える環境の充実が望まれる所である。

以上、示してきた各分野が抱える課題は、いずれも統計手法を利用できる環境が十分でないことが原因の一端となっている。統計手法の利用環境がもっと普及すれば人材の育成をはじめとして多くの課題の解決に向けて前進する事ができると考えられる。

このような環境の整備の一助になればと考えて開発したのが StatPartner である。

3. StatPartner が目指すものとその特徴

これまで述べてきた状況下で必要になる統計解析ソフトウェアが備えるべき条件をまとめると次のようになる（吉田・石橋（1995））。

- 必要な統計手法が利用できること
- コンピュータや統計解析に不慣れな人、いわゆる入門者が少ない学習時間で操作できること
- 一般にある環境で動作できること
- 導入コストが低いこと

StatPartner はこのような条件を満たすことを目標に開発を行った。そして、前章で述べた統計手法の活用現場で有効に利用でき、統計手法が活用されるフィールドが広がることを目指している。

StatPartner の特徴を以下に示す。

- 豊富な解析手法を持つ
- 操作が Windows に準拠しており容易に使い始められる
- 入出力画面は完全に日本語化されており、データにも日本語の文字列が利用できる
- 解析を行いながら統計量と手法を解説した「知識」を参照できる
- 価格をワープロソフトのキャンペーン価格程度に押さえている
- 標準的なハードウェア構成で動作するため、利用時に新たな投資が不要である

つまり、手軽に「手に入れることができ」、「解析する事ができ」、「不明点を解決できる」入門者向けの統計解析ソフトウェアである。

4. StatPartner の歴史・動作環境・機能

StatPartner は、ある大学の新生教育用計算機に添付する目的で開発を行った。そして 1995 年 7 月から基本統計機能のリリースを開始し、その後、多変量解析のオプションなどをリリースし現在に至っている。ソフトウェアの体系と動作環境を表 1~3 に示す。

5. StatPartner の利用例

(1) 解析

データは図 1 に示すデータシートに入力する。データ値は連続量と離散量の 2 種類で管理される。変数名、サンプル名、離散量のデータ値に日本語の文字列が使えるため初心者の理解に役立つ。

図 2 はメニューバーから分析対象の変数名を指定するダイアログボックスを呼出し、変数を

表1. ソフトウェア体系.

ソフトウェア名	内容	リリース時期
StatPartner BASE	基本的な統計とグラフ、データ管理、データ操作	1995年7月
StatPartner 多変量解析パック	BASE に多変量解析を追加	1995年10月
StatPartner 検定パック	BASE にノンパラメトリック検定と多重比較を追加	1996年4月
StatPartner フルパック	BASE に多変量解析、ノンパラメトリック検定と多重比較を追加	1996年4月

表2. 動作環境.

動作OS	Windows3.1以降 (WindowsNT3.1以降、Windows95 動作可)
動作機種	i486以上のCPUを搭載したAT互換機またはPC98シリーズ
メモリ	8MB以上 (WindowsNT、Windows95の場合は16MB以上)
ハードディスク	7MB以上

表3. 機能一覧.

総合スペック		サンプル数 最大 1990、変数 最大 390	
BASE	統計手法	基本統計、相関分析、一元配置/二元配置分散分析、推定・検定 (平均値、平均値の差、分散、分散比、分割表)、クロス集計	
	グラフ	ヒストグラム、散布図、棒グラフ、折れ線、時系列のあてはめ、樹葉図、箱ヒゲ図、正規確率プロット	
	データ管理	変数・サンプルの挿入・削除、変数変換、サンプルの抽出、ソート、連続量の離散量化、カテゴリの統合、他ソフトとのデータ交換 (CSV形式、TAB形式、クリップボード経由)	
多変量解析パック		重回帰分析、線形判別分析、正準判別分析、主成分分析、因子分析、クラスタ分析、数量化理論1類、2類、3類	
検定パック	ノンパラメトリック分析	1標本	Kolmogorov-Smirnov
		2標本	符号検定、Wilcoxon 検定、Kolmogorov-Smirnov、Mann-Whitney のU検定、Fisherの直接確率検定、中央値検定
		k標本	Friedman 検定、中央値検定、Kruskal-Wallis 検定
	順位相関係数	Spearman の順位相関係数、Kendall の順位相関係数	
多重比較		Scheffe、Bonferroni、Tukey、Duncan、Dunnnett	

X1[連続量]	X2[連続量]	X3[連続量]	X4[連続量]	X5[連続量]
蛋白質(g)	脂質(g)	炭水化物(g)	炭水化物(g)	灰
S1 玉露	29.100	4.100	32.700	11.100
S2 抹茶	30.700	5.300	28.600	10.000
S3 煎茶	24.000	4.600	35.200	10.600
S4 かまいり	24.200	3.500	35.600	10.700
S5 番茶	19.700	4.400	33.500	19.500
S6 ほうじ茶	18.200	4.800	39.200	18.700
S7 玄米茶	13.000	3.400	62.900	7.900
S8 ウーロン	19.400	2.800	39.800	12.400
S9 紅茶	20.600	2.500	32.100	10.900

図1. データシート。

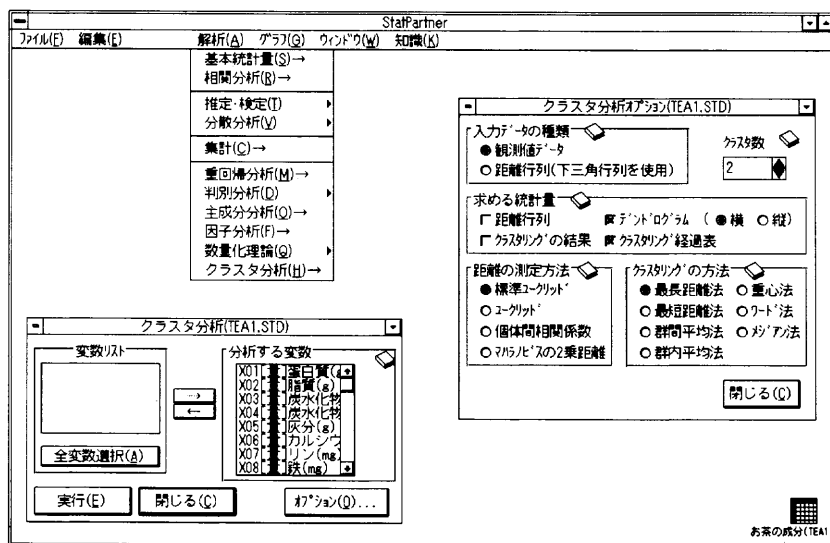


図2. 分析パラメータ指定ウィンドウ。

指定した所である。ここで実行ボタンをクリックすると分析結果が表示される。必要に応じて分析オプションを指定する事もできる。

図3が求めた分析結果である。分析結果は新たなデータシートとグラフに出力される。データシートに出力された分析結果はグラフを描くなど、次の分析の対象データにする事ができる。つまり残差を求め、その残差が格納されたシートを元に棒グラフや折れ線グラフを表示することができる。

StatPartnerの分析機能とグラフ表示機能はこの様にシンプルになっており、同一の手順で操作する事が出来る。従って、ほとんどパソコンにさわたた事のない学生でも15分程度あれば目的とする結果を求めることができる。

(2) 知識

統計手法と用語を解説した知識はメニューバーや分析結果のブックアイコンをクリックする

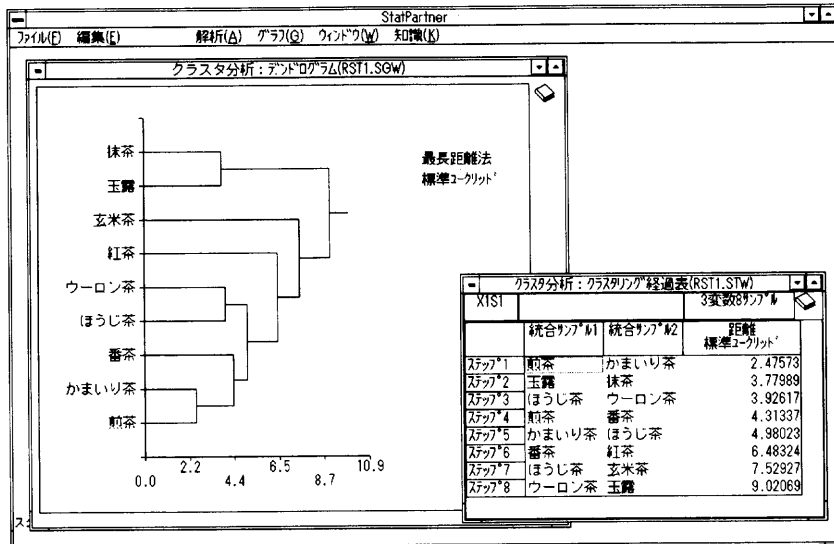


図3. 分析結果.

統計手法のチュートリアル

重回帰分析とは

- 目的変数
- 説明変数
- 回帰係数
- 分散分析表
- 適合度
- 残差
- 影響度
- 変数選択の要約
- 変数追加法
- 変数減少法
- 変数選択法
- 基準値

重回帰分析(重回帰分析--概要5)

経を単純にするために、ここでは説明変数(要因)が1つの場合で説明します。
 目的変数: 県民所得 (y)
 説明変数: 製造業の比率 (x)
 とします。
 2つの変数による散布図をみると、なにか連続的な関係がありそ

県民所得 y
製造業 x

解析のダイアログ

重回帰分析(SCORE310)

目的変数: 県民所得
説明変数: 製造業の比率

分析結果

XIST	変数	自由度	不偏分散	F値	p値	検定
回帰変数	5279.38924	2	2638.19200	31.79516	0.00000	***
誤差変数	3974.11750	37	83.08424			
全変数	8247.50000	39				

統計用語の解説

分散分析表

求められた回帰式全体が有意かどうかの検定を行うことができます。帰無仮説は、すべての回帰係数=0であり、もし帰無仮説が棄却されなければ、求められた回帰式は意味をなさないと考えられます。このような場合はもともたデータ自体を見直す必要があります。

データ全体の分散(全変動)をつぎのように分解することができます。
 全変動 = 回帰変動 + 誤差変動
 これらの値をもとにして、F値とそれに対応するp値を計算します。p値は予め定められた値(0.05または、0.01)と比較さ

図4. 知識の表示例.

事により適時呼出すことができる(図4)。

このように StatPartner は手軽に解析結果を求めることができるため、大学だけでなく企業などでも数多くご利用いただいている。大学ではパソコン教室、学生保有パソコンに加え卒業

研究や一般の研究に、また企業では企業内教育をはじめとして、リスク分析等にもご利用いただいている。

6. チェック&アドバイスシステム (StatGuide)

StatPartner は統計解析を利用できる環境を提供するが、今回の対象である統計手法を使い始めた人たちに焦点を当てた場合、さらに取り入れるべき機能や考え方が幾つかある。その一つが解析の過程をシステム側でチェックし、必要なアドバイスを利用者に与える機能である。

出力された統計量をシステム側でチェックし解析者にアドバイスを与える事により、より適切な解析結果を導く事ができる。このような考え方は古くからありその試作システムも幾つか作られた。仕組みとしてはプロダクションルールや知識をクラス化したエキスパートシステムなども試みられ、現在も研究が続けられている。

StatPartner の関連ソフトウェアにもこれに類する機能を持つ StatGuide for StatPartner (以下 StatGuide と称する) がある。StatGuide は StatPartner に類似したインターフェースを持ち、解析途中で必要な統計量のチェックを行い、必要なアドバイスを与えるものである。このような機能を持つ関係上 StatPartner ほどの柔軟性はないが、StatPartner だけでは見逃す可能性のあるチェックが自動化されているため StatPartner と相互に補完した形で利用する事により、解析効率や教育効果を上げることができる。なおアドバイスを提示する仕組みは、統計量をチェックするのに必要な知識をチェックシートとして定義し(赤澤 他(1995))、これに基づきチェックを行いアドバイスを与える形式である。以下に StatGuide の概要を紹介する。

(1) StatGuide の主な機能

StatGuide は多変量解析のなかでも予測・判別に用いられる、重回帰分析、線形判別分析、正準判別分析、数量化理論(数量化1類, 数量化2類)の5手法についてサポートする。動作環境等はほぼ StatPartner と同様である。主な機能を下記に示す。

- 手法の自動選択機能
データから解析に用いる手法を自動的に選択する機能。
- 分析の前のデータの事前解析, 分析結果の解析評価機能
岡山大学との共同研究による, 統計解析手順チェックシートをシステム化。正しい統計解析の手順に沿って, 適切なアドバイスを利用者に与える。
- 分析の実行手順の履歴保存機能
実際に行った解析の手順, チェックの状況を履歴として後で参照することが可能。

(2) 重回帰分析による StatGuide の利用例

StatGuide は解析を3段階で行う。まず、分析を行うためにデータが解析に適しているかどうかを検討する「事前解析」、次に解析結果を検討する「解析評価」、最後に解析を行った後の「事後解析」を途中の統計量をチェックしながら進めていく。

また、StatGuide の実行モードには2つのモードがある。「Step モード」は事前解析, 解析評価, 事後解析をワンステップずつ結果を見ながら実行し、「Auto モード」は自動的にチェックを行いながら, 最終結果まで自動的に実行する。以下の例は Step モードによるものである。

メニューバーの実行から「分析手法が不明で実行」を選択すると図5が表示される。

パラメータとして、予測する変数と、予測するための変数を指定し判定を行うと図6が表示され、このデータが重回帰分析に適していることが分る。

分析判定ウィンドウから「次へ」を実行すると、重回帰分析を3段階でチェックしながら実

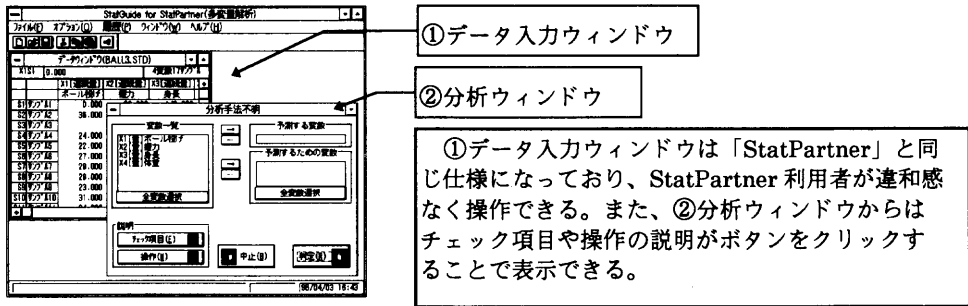


図5. StatGuide手法の自動選択.

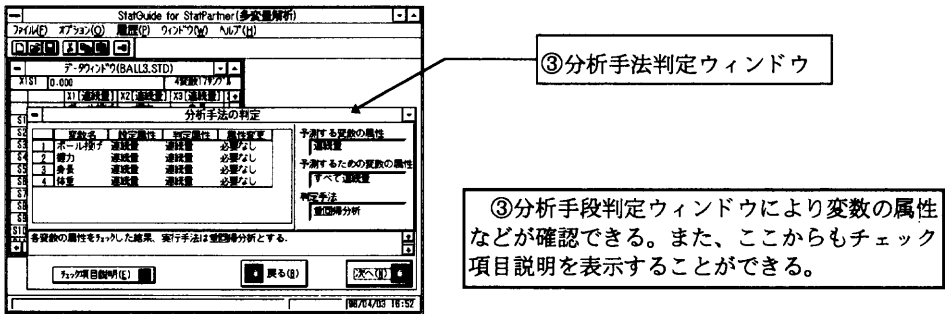


図6. StatGuide分析手法の判定.

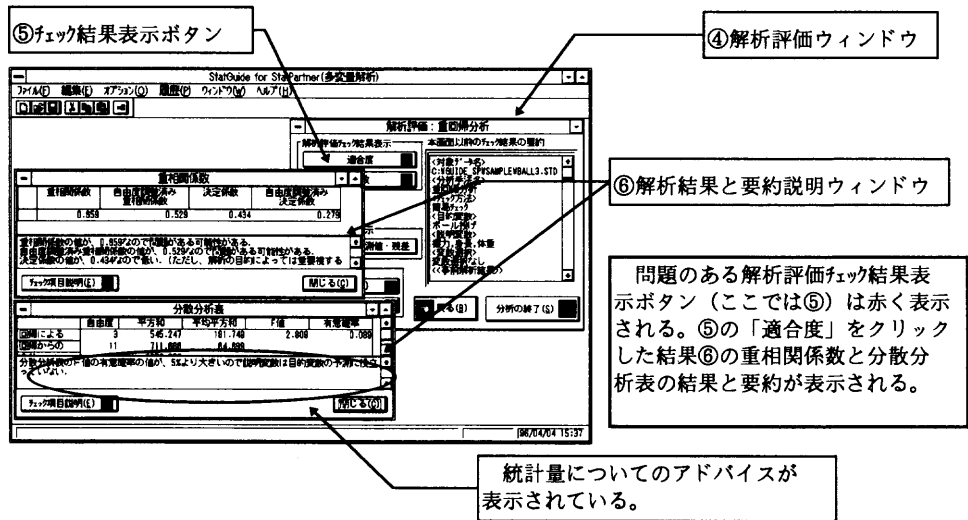


図7. StatGuide解析評価.

行する。

図7は、解析評価に問題があったため、適合度のチェック結果を表示したものである。この他 StatGuide には解析履歴を残し、後で手順をトレースできる機能やチェックのレベル

を変更できる機能など、数々の便利な機能が利用できるようになっている。このように Stat-Guide はデータ解析の経験が少ない人でも、適宜、効果的なアドバイスを受けながら解析が行える仕様になっている。

7. 最 後 に

多変量解析は計算機の利用を前提とした手法であり、従って計算機とともに発展・普及してきたと言える。StatPartner は Windows の提供する環境や Visual Basic などのエンドユーザコンピューティングツールなしには機能面・コスト面などで実現しなかった製品である。インターネットやパソコンの低価格化がこの分野に何をもたらすのであろうか。データの解析をする局面がますます多くの層に広がり、それぞれが自分なりの解釈を持てる時代が来ることを期待する。そしてそのために、今後も StatPartner などに有益な考え方や機能を取込み、進歩が著しい情報処理環境を有効に活用できる仕組み作りに取り組んで行く予定である。

参 考 文 献

- 赤澤利夫, 井上恭子, 垂水共之 (1995). チェックシートによる統計解析の知識表現 (3)—プロトタイプ
の作成と評価—, 第9回日本計算機統計学会発表論文集, 44-47.
- 吉田典弘, 石橋雄一 (1995). 新統計教育のソフトウェア (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ウム論文集, 63-72.

Development of StatPartner and Its Application

Norihiro Yoshida and Shigeto Kudo

(NEC Software Ltd.)

With the expansion of information technology such as personal computer and the Internet, it becomes general to apply statistical methods, including multivariate analysis, to various kinds of investigations. The present condition of statistics for beginners and usage of statistical methods are surveyed in this article. Motivation of development of StatPartner and specification are explained.