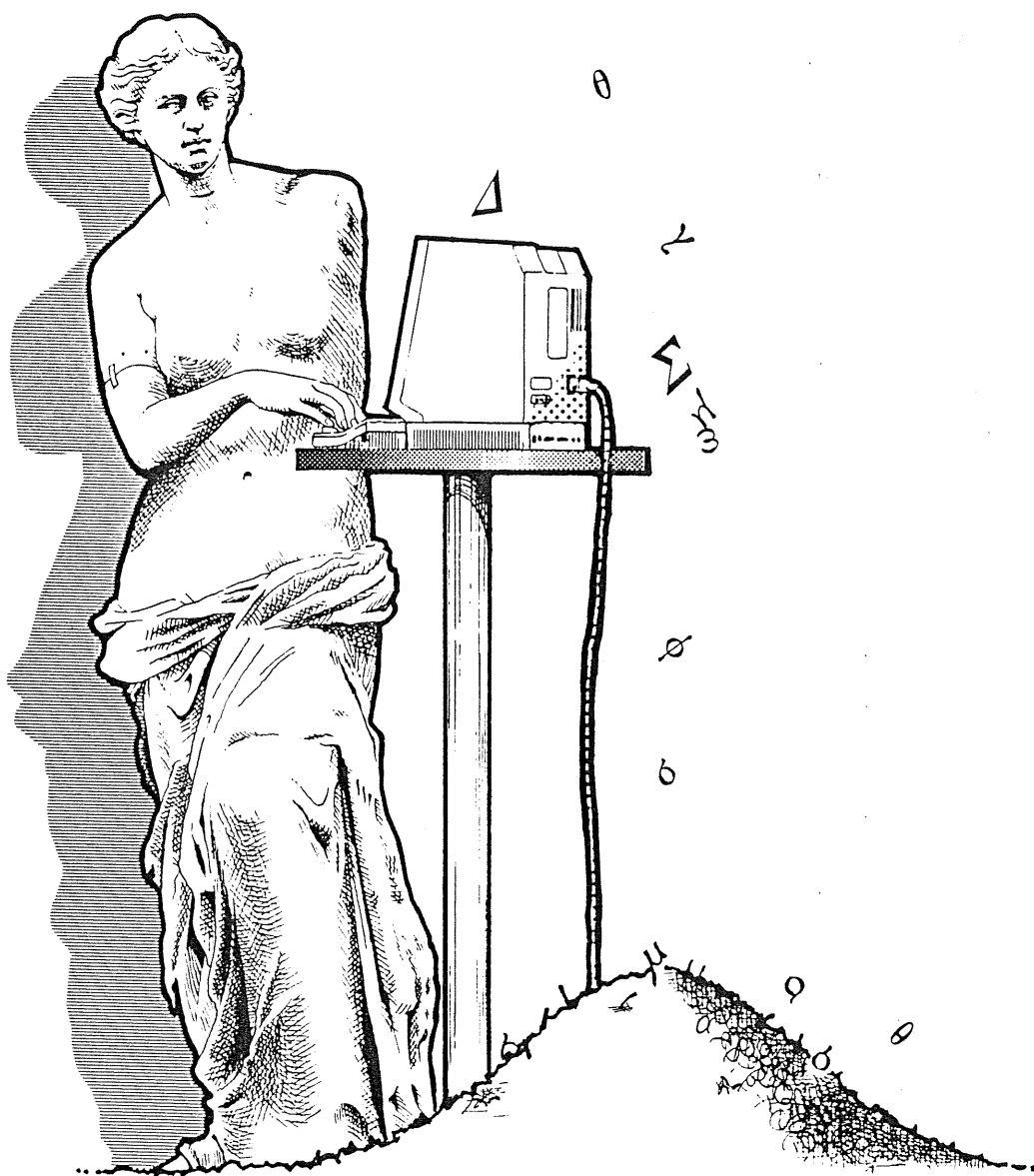


第16回日本行動計量学会大会 発表論文抄録集



昭和63年 8 月25日(木)～27日(土)

於 千葉大学文学部

統計学パッケージ HALBAU の開発について

○高木廣文^{*1}, 柳井晴夫^{*2}, 市川雅教^{*3}, 服部芳明^{*4},

佐藤俊哉^{*5}, 丸井英二^{*5}, 佐伯圭一郎^{*5}, 中井里史^{*5}

*1 聖路加看護大学看護学部 *2 大学入試センター研究開発部 *3 東京工業大学工学部

*4 日本原子力事業(株) *5 東京大学医学部

【はじめに】

いわゆる「パソコン」の普及とその高性能化に伴い、大型計算機でなければ計算が不可能であった多変量解析の手法の多くが、演算時間はかかるもののパソコン上で実行可能となった。また、パソコンではグラフ表示が比較的簡単なプログラムで可能であり、データの視覚化を容易にしてくれる。さらに、今まで大型計算機を手軽に利用できない環境にあった研究者にとって、パソコンでデータ解析が行えることは、極めて望ましいことと考えられる。しかし、現実の調査データでは質問項目数が100を越えるものが少なくなく、またケース数も1000例以上あるものは珍しくない。従って、データ解析を行うためには、パソコン用のソフトといえども、ある程度の制限はあっても、実用的な水準を保つ必要があるだろう。また、調査票から解析可能なデータファイルとして、ディスクなどの媒体に保存したり、データの訂正や編集を行うためのエディタも完備されている必要があるだろう。また、分析結果を単に画面上やプリンタから出力するだけでなく、現在普及している日本語ワープロなどで処理できるように、テキストファイルとしてディスク上に出力できることも必要な機能であると考えられる。

HALBAU (High quality Analysis Libraries for Business and Academic Users) は、上記の要求を満たし、実際の調査研究に使用可能なデータ解析用ソフトとして開発されたものである。調査データの入力・編集から各種の分析結果の出力までを、メニュー画面への簡単な応答により、複雑な操作を必要とせずに実行できるように設計されている。ここでは、HALBAUのシステム構成や機能などについて報告を行う。

【HALBAUの構成内容】

使用言語は NEC-N88日本語BASIC (86) (MS-DOS版) であり、NEC PC9800シリーズの全機種と 98XA, XL, XL²上のインタープリタ、もしくはコンパイル後 MS-DOS 上で実行可能である。

表に示したように、HALBAUは大きく3つの部分から構成されている。それらは、(1)データファイルの操作関係パッケージ、(2)基礎統計学パッケージ、および(3)多変量解析パッケージである。ただし、各解析方法の分類は便宜的なものであり、必ずしも最適な分類を行っている訳ではない。

(1)のデータファイルの操作関係パッケージは、パソコンの操作やデータ解析の経験の少ないものでも、簡単に手持ちの調査票などからデータを入力し、ディスク上にデータファイルとして保存できることを意図して作成されたものである。さらに、データの誤入力などを訂正し、編集するのも比較的簡単に行えるように設計されている。また、複数個のデータファイルを統合するために、ファイルのチェインやマージが可能となっている。実際の調査では複数個の項目のデータから、新しい尺度を構成することがよく行われるが、この目的のために、簡単なデータ変容ルーチンを用意した。ただし、複雑なデータ変容を行うためには、変容プログラムをエディタ上で自作する必要がある。BASIC 言語に関する知識のない、全くの初心者には使用は困難であろう。なお、データは付属のエディタにより、最大約5000変数まで入力可能であり、ケース数に制限はない。しかし、ハードディスクなどが使用できない場合には、ケース数や変数の個数の増加につれ、入出力時間はかなり長くなる。また、多変量解析等で一度に分析可能な変数の個数は約80程度である。多くの手法では分析できるケース数に制限はないが、最尤解や順位を求めるような分析では、(ケース数)×(変数の個数)が16000以下となる必要がある。また、他の形

表 HALBAUの構成内容

(1)データ・ファイルの操作関係
①データ入力・編集・出力等 ②他形式ファイルとの相互変換 ③簡易型データ変容等
(2)基礎統計学パッケージ
①基本統計量の計算 ②項目別カテゴリ度数の算出 ③度数分布表とヒストグラムの作成 ④幹葉表示の作成 ⑤クロス表の作成と検定 ⑥相関係数の計算と検定 ⑦相関図の作成(＋単回帰分析) ⑧2群の母平均値の差の検定 ⑨一元配置分散分析(＋対比較) ⑩四分表の併合 ⑪3変数以上の同時相関図の作成 ⑫順位和検定, 符号検定, 符号付 順位和検定 ⑬Kruskal-Wallis検定(＋対比較) ⑭二元配置分散分析(＋対比較) ⑮5数要約値と箱ヒゲ図
(3)多変量解析パッケージ
①重回帰分析と数量化理論1類 ②主成分分析 ③正準相関分析と数量化理論3類 ④判別分析 ⑤数量化理論2類 ⑥非線形モデルと生命表解析 ⑦因子分析 ⑧クラスタ分析

式のデータファイルとの相互変換は dBASE-II, MULTIPLAN(SYLK形式) および固定長データファイルとの間で可能である。

(2)の基礎統計学パッケージは、分布の特性値である平均値、分散、尖度、歪度などを求めたり、度数分布表やヒストグラム、幹葉表示、箱ヒゲ図などを作成し、主にデータの特性を記述し、探索的データ解析がある程度可能になっている。また、通常用いられている検定手法に関しても、クロス表の検定やノンパラメトリックな方法と共に、各種の手法が組み込まれている。

(3)の多変量解析パッケージは、より複雑な事象間の関係进行分析するために作成されたものである。「予測」のために、重回帰分析と数量化理論1類があり、「識別問題」のためには判別分析と数量化理論2類がある。また、「内的構造分析」のためには因子分析と数量化理論3類があり、「情報圧縮」のためには主成分分析がある。また、変数群の関係を分析するためには正準相関分析と双対尺度法(数量化理論3類)がある。「個体分類」や「項目分類」のためにはクラスタ分析があり、「生命表解析」のためには比例ハザードモデルおよび Kaplan-Meier 法、一般化ウィルコクソン検定等が組み込まれている。

なお、各手法の分析結果はグラフ表示以外は、プリンタかディスク上のファイルに出力することができる。

【問題点】

実際の使用状況からみて、データ解析の初心者でも、独力でかなり高度な分析が行えるようである。また、一度作成したデータファイルは、全ての分析手法で共通に使用できるため、入力に時間を要してもその後の処理が簡単にできるため、分析手法にそれほど詳しくなくとも、一通りの分析結果を得ることができるようである。しかし、実際にはこの点が最も重要な問題点でもある。すなわち、HALBAUの使用は、統計的解析の誤用を助長する可能性が極めて大きいともいえる。含まれる手法は多数あり、必ずしもデータに適した解析方法を用いるという保証はない。さらに、結果の解釈を適切に行えるのかという疑問もある。実際にユーザからの問い合わせは、大部分が方法の誤用に起因するエラーについてである。この対策をどのようにプログラムに組み込むかが当面の最大問題である。ただし、この問題は一般に全ての統計学ソフトで共通することである。データ解析の完璧なエキスパートシステムが完成するまでは、ユーザ自身が、データ解析の各手法を正確に理解し、解析結果の解釈に必要とされる幅広い教養、専門知識を身に付け、さらに直観力、知恵、知力を十分育成する以外に問題解決の方法はないと考えられる。

【参考文献】

柳井、高木 編著：多変量解析ハンドブック，現代数学社，1986