計算機統計学
第5巻・第2号：1992
165−169

統計パッケージHALBAUについて*

高木廣文**

要　旨

調査資料に基づく分析を，手元にあるパーソナル・コンピュータ（パソコン）で簡単に行えれば，極めて便利である．最近の技術的進歩により，パソコンの性能も著しく向上した．その結果，以前では考えられなかったような，高度な統計計算が手軽に行えるようになった．我々は，そのような背景のもとで，データ解析用パッケージHALBAU（High quality Analysis Libraries for Business and Academic Users，通称「はるぼう」）を開発した．ここでは，そのバージョン3.33版について紹介する．

1. HALBAUとは

HALBAUは，『多変量解析ハンドブック（柳井・高木，1986）』に収録されたBASIC言語による多変量解析用のサブプログラムを主要計算ルーチンとして開発され，㈱現代数学社から販売されている．ただし，同書に掲載されているプログラムリストを，単にディスクに移しただけのものではない．データ入力・編集，分析方法の指定のためのメニュー，分析のための変数指定および結果の出力用のプログラムなどを加え，システムとしてより使いやすいものとなっている．また，実際のデータ解析では，いきなり多変量解析の方法を用いるのは，あまり感心すべき手順ではない．始めは単純な分析により，データの点検やその特性を探索するのが普通であろう．そのため，基礎統計学の基本的な手法が大幅に，HALBAUにも加えられている．さらに，分析のためにデータを適当に加工することが必要であれば，HALBAU内にある変容ルーチンを用いて，比較的簡単に行えるようにも注意が払われている．このように，HALBAUはデータ入力から分析法の指定，実際の分析，結果のプリンタ出力・ファイル出力に至るまでを，統計学やデータ解析の初心者でも，比較的容易に行えるように作成された，統合型の統計学プログラム・パッケージである．

2. HALBAUの特徴

HALBAUの開発で留意した点は，初心者でも簡単に扱え，かつ実用的な統計パッケージを提供するということであった．とくに，大型計算機やワークステーションがなくても，パソコンさえあれば，実際にデータ解析をかなりの程度まで行えるようにすることであった．その結果，HALBAUは次のような特徴をもつことになった．

(1)メニュー画面方式を採用している．
(2)データ・エディタなどを装備している．
(3)欠損値の処理が可能である．
(4)日本語による指示・表示が行える．
(5)結果のファイル出力が可能である．
(6)データ変容がある程度まで可能である．
(7)豊富な分析手法を含む．

---

* *Statistical package HALBAU* by Hirofumi Takagi (The Institute of Statistical Mathematics, Tokyo 106, Japan).
** 統計数理研究所　〒106 東京都港区南麻布4−6−7（☎03−3664−1695）

(8)他の形式のデータ・ファイルと相互互換性をもつ.

(9)大標本，および多項目データの処理が可能である.

3. HALBAUの構成

HALBAUは図1に示すように，大きく三つの部分から構成されている．すなわち，「1. データファイルの編集と変容」,「2. 基礎統計パッケージ」，および「3. 多変量解析パッケージ」である.

表1に示したように，データ・ファイルの編集と変容のパッケージは，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．また，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．さらに，複数個のデータ・ファイルを統合するために，ファイルのチェー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も用意されている．ただし，複雑なデータ変容を行うためには，そのためのプログラムをエディタ上で自作することが必要であり，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る．なお，データは付属のエディタにより，最大約5000個の変数まで入力可能であり，基本的に作成できるデータ・ファイルについてのケース数の制限はない．また，他の形式のデータ・ファイルとの相互変換はdBASEⅡ，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.

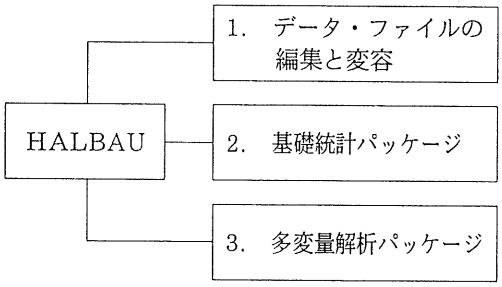

図1　HALBAUの構成

表2に示したように，基礎統計に関するパッケージでは，単純集計や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できるようになっている．これは，主にデータの特性を記述し，ある程度の探索的データ解析ができるように意図されたためである．また，通常よく用いられる検定方法に関しても，母平均値の差の検定，クロス表の検定，分散分析などとともに，ノンパラメトリック法も組み込まれている.

表3に示した多変量解析パッケージは，より複雑な事象間の関係を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Ⅰ類があり，「識別問題」のためには判別

表1　データおよびファイルの操作

名　称	機　能　説　明
データ・ファイルの作成および編集	データ・ファイルの作成と編集：データの入力・編集，ディスクへの保存，変数の削除と追加など．変数は約5000個以下. チェーンとマージ：二つのファイルの一つのファイルへの統合. データの変容：変容プログラム作成用エディタとその実行. プリンタ出力：HALBAU用ファイルの清書出力. ファイルの消去など.
他の形式のデータ・ファイルとの相互変換	HALBAU用データ・ファイルと固定長データ・ファイル，MULTIPLANのSYLK形式ファイル，およびdBASEⅡファイルとの相互変換.
データの簡易型変容とサブファイルの作成	複数個の変数の重み付き・ベキ乗付きの線形和の計算，変数の再カテゴリー化，変数の対数変換，HALBAU用ファイルの分割など. 特定の変数の値によるファイルのソート・マージ.

表2　基礎統計パッケージ

名　称	機　能　説　明
基本統計量の計算	平均値，（標本）分散，標準偏差，変動係数，歪度，尖度など.
項目別カテゴリー度数	質的変数の各カテゴリーの度数と百分率
度数分布表とヒストグラムの作成	適当な階級区分での度数，累積度数，相対度数，累積相対度数，ヒストグラム，折れ線グラフなど.
幹葉表示の作成	適当な階級区分での幹葉表示.
クロス表の作成と検定	一般のクロス表の場合：カイ2乗値，自由度，有意確率，関連係数．四分表の場合：連続性の補正を行った場合と行わなかった場合のカイ2乗値，有意確率，Fisherの正確な確率，リスク比とオッズ比とその信頼区間など.
四分表の併合	層別四分表についてのMantel−Haenszel推定量. コホート研究の場合：各層でのリスク比，リスク差，Cochran−Mantel−Haenszel検定(カイ2乗値，有意確率)，共通リスク比とその信頼区間など．ケース・コントロール研究の場合：各層でのオッズ比，Cochran−Mantel−Haenszel検定(カイ2乗値，有意確率)，共通オッズ比とその信頼区間など．断面研究の場合：各層でのYuleのQ，ファイ係数，Cochran−Mantel−Haenszel検定(カイ2乗値，有意確率)，共通YuleのQなど.
相関係数の計算と検定	平均値，標準偏差，変数間のPearsonの積率相関係数，無相関性の検定（t値，自由度，有意確率).
相関図の作成	平均値，標準偏差，変数間のPearsonの積率相関系数，無相関性の検定（t値，自由度，有意確率)，単回帰式と原点を通る回帰式，相関図（散布図).
同時相関図の作成	3変数以上の相関係数行列のわかる同時相関図.
2群の母平均値の差の検定	独立2標本の場合：各群の平均値，分散，分散比(F値)，自由度，有意確率（等分散性の検定)，等分散の場合の2群の母平均値の差の検定(t値，自由度，有意確率)，不等分散の場合のWelchの検定．対応のある場合の母平均値の差の検定.
1元配置分散分析	各群の平均値，分散，級間変動，級内変動，全変動およびその不偏分散，分散比（F値)，自由度，有意確率など.
順位和検定，符号付き順位和検定など	Wilcoxonの順位和検定（U検定)．対応のある場合には，符号検定，Wilcoxonの符号付き順位和検定など.
Kruskal−Wallis検定	Kruskal−Wallis検定と対比較（Bonferroni，Scheffeの方法).
2元配置分散分析	各要因の水準ごとの平均値，（不偏）標準偏差，各要因変動，交互作用，残差変動，全変動およびその不偏分散，分散比（F値)，自由度，有意確率．交互作用を残差に含めて再計算が可能など.
5数要約値と箱ヒゲ図	最小値，第1四分位，中央値，第3四分位，最大値，四分位偏差，箱ヒゲ図など.
散布図の作成	層別変数の値ごとの散布図.

分析と数量化理論Ⅱ類がある．また，「内的構造分析」と「情報圧縮」のためには主成分分析，因子分析，数量化理論Ⅲ類がある．また，変数群の関係を分析するためには正準相関分析などがある．「個体分類」や「項目分類」のためにはクラスター分析があり，「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Wilcoxon検定などが組み込まれている．

表3　多変量解析パッケージ

名　称	機　能　説　明
重回帰分析と 数量化理論Ⅰ類	重回帰分析の場合：各変数の平均値，標準偏差，変数間の単相関係数，偏回帰係数，標準偏回帰係数，F値，有意確率，重相関係数，多重決定係数と検定(F値，自由度，有意確率)，自由度調整重相関係数，自由度再調整重相関係数．説明変数の選択方法として，一括投入法、変数増加法・変数減少法，変数増減法，変数減増法（F値，AIC基準）．残差プロットなど．数量化理論Ⅰ類の場合：説明変数間のクロス表，カテゴリーごとの基準変数の平均値と標準偏差，標準化数量，基準変数との偏相関係数，重相関係数など．
主成分分析	各変数の平均値，分散，標準偏差，変数間の相関係数．主成分負荷量，固有値とその検定，寄与率，累積寄与率，主成分得点．
正準相関分析と 数量化理論Ⅲ類	正準相関分析では，各変数の平均値，標準偏差と変数間の相関係数，各変数の重みベクトル，構造ベクトル，各成分の固有値の検定，正準相関係数，冗長性など．数量化理論Ⅲ類では，各カテゴリーおよびケースの重み数量，各成分の固有値など．
判別分析	2群および多群の判別．グループ別に各変数の平均値，標準偏差，変数間の相関係数，各変数の判別係数，Bartlettのカイ2乗検定，自由度，有意確率，BoxのM検定（カイ2乗値，自由度，有意確率)，各群の判別得点の平均値，誤判別率など．説明変数の選択として，一括投入および変数増加法．
数量化理論Ⅱ類	2群および多群の判別．群ごとの説明変数間の各カテゴリーのクロス表，カテゴリー数量，群ごとの判別得点の平均値など．
非線形モデルと 生命表解析	多重ロジスティック・モデル，多重 Weibull，比例ハザード・モデル，Kaplan-Meierの生命表解析，一般化Wilcoxon検定．各手法に関して，各モデルに含まれる説明変数の係数，F値，有意確率，AICなど．三つの生命表解析の手法に関して，期間別生存率・累積生存率など．累積生存率のプロット．
因子分析	主因子法，最尤法．共通性の初期値指定，反復推定の指定可能．因子の回転として，直交回転(バリマックス法，バイコーティマックス法，コーティマックス法)，斜交回転(コバリミン法，バイコーティミン法，コーティミン法)．回転前・後の因子負荷量．斜交回転の場合には，回転後の構造行列，因子間の相関係数行列．因子得点の指定として，残差平方和最小，独自性を除外，Thomsonの方法，Bartlettの方法など．
クラスター分析	個体分類の場合，データを標準化するか，そのままか，もしくはMahalanobisの汎距離かの指定可能．項目分類の場合，相関係数から距離に変換．分析手法として階層的手法（最近隣法，最遠隣法，メディアン法，群平均法，重心法，Ward法)．デンドログラムの出力．

4. HALBAUの制限について

実際の分析を行う場合，どの程度の規模のデータが扱えるのか，関心のあるところであろう．

データ・ファイルを作成する場合，変数の個数は最大で約5000まで，ケース数の制限はないが，最大でフロッピィ・ディスクやハード・ディスクなどの記録媒体の容量までとなる．ただし，データ・ファイルの作成時に作業用の領域を確保するので，作業用と保存用のディスクを別に指定しないと，容量いっぱいにファイルを作成することはできない．また，ケース数を整数型で数えているので，32767ケースを越えるとエラーになる．実際に，パソコンでそれ以上のケース数のデータを扱うのは，やや無謀ではあるが，HALBAUで若干プログラムを修正すれば可能である．

しかし，実際の分析では，一度に使用できる変数の個数に関して制限がある．これは，BASICプログラムにおける一つの配列が確保できる記憶

領域の制限と，使用機種の記憶容量に関する制限によるものである．BASIC言語の制限から，一度に分析できる変数の個数は最大で89個となる（多変量解析の場合）．

分析できるケース数の制限は，とくに設けていない．しかし，「数量化理論Ⅲ類」では，ケース数とカテゴリー数の積が約32000以下，「非線形モデルと生命表解析」では，ケース数と変数の個数の積が約16000以下までの場合にしか分析できない．ただし，この場合でもカテゴリー数や変数の個数は89以下でなければならない．また，基礎統計パッケージの「順位和検定など」および「5数要約値と箱ヒゲ図」には「非線形モデルと生命表解析」とほぼ同様の制限がある．「クラスター分析」では，変数の個数は125個以下となる．ただし，「項目分類」の場合はケース数に制限はないが，「個体分類」の場合は125ケース以下しか扱えない．

5. HALBAUの問題点と今後

実際に自分で使用し，また学生・院生，他の研究者などからの意見を聴くと，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．ただし，今後の主な課題と問題点として

(1)メニュー画面方式なので，指定がやや面倒である，

(2)グラフィクスによる表示が少ない（貧弱である），

(3)デフォルト処理があるため，訳のわからない分析が行われる，

(4)より多様な分析方法が必要である

などがある．

(2)と(4)については，徐々にHALBAUに組み込んでいく予定ではある．(1)と(3)は相互に関係する問題である．各種の分析のための指定について，手間をかけないようにするには，デフォルト処理を増やさねばならない．あまり，デフォルト処理を増やすと，自分が何を行っているのか理解せずに，分析してしまう．場合によっては，得られた結果の解釈が不適切な，さらには誤った結論を導くことにもなる．我々も，各分析手法の何をデフォルトにするかを，慎重に決めることが必要である．しかし，使用者もデータ解析の各手法をある程度まで理解してから，分析にあたるべきであろう．

現在，HALBAUの改訂作業を行っているが，これらの問題の一部でも解決できればと考えている．なお，HALBAUについて，より詳細な情報が欲しい場合には，下記の文献を参照していただきたい．

参考文献

柳井晴夫・高木廣文編著（1986）．多変量解析ハンドブック．現代数学社．

高木廣文・佐伯圭一郎・中井里史（1989）．HALBAUによるデータ解析入門．現代数学社．