STATISTICIAN'S APPRENTICE:
AN EXPECTED STATISTICAL PACKAGE OF THE NEXT GENERATION

Joji KARIYA

Information Processing Center, Yamaguchi University

SUMMARY

Current famous statistical packages -- PRODAS, P-STAT, S, SAS and our SPMS -- are investigated and compared with each other. It is revealed that they have little clear-cut conception of data and relationship among them. Facilities and structures of a next generation statistical package are described. In an environment offered by a new statistical package on a personal computer, everyone can participate in creation and exchange of knowledge, including those of a system itself. The most important component of a system, called a statistician's apprentice, is an agent of a user. It provides a user with his personal environment, and intermediates between him and specialists in the system.

1. まえがき

統計パッケージとして何が望ましいかは難しい問題である。可用性（使用出来る機能が多い）、効率（計算が早い）、理解性（思考のレベルが利用者とあっている）、その他色々な評価基準があり、立場により異なる。特に統計家は利用者と開発者の2面で関与するので複雑である。統計パッケージは、新たな知識を吸収して常に進化すべきものであり、発展性を特に重視したい。

我々は10年前から統計パッケージSPMSの開発を始め、それを実用レベルに到達させるのに4年を要した。SPMSはある意味では成功し、ある意味では失敗であった。すなわち、SPMSを用いると、構造的にも、量的にも、多種のデータについて解析を行うことができ、十分な能力を示した点では成功であった。しかし、ハードウェアの能力不足を補うためのSPMS固有の命令を使用する記述を含むプログラムが多いため、現在使用できる、より強力な計算機を使用しようとする、移植性が悪く、また移植出来ても、計算機の能力を生かせない点では失敗であった。さらに重要な事は、データ構造がファイル上の表現と密接に対応しているため、概念的拡張が困難なことである。

私は今、REWMOL (REal World Modelling Language) という名前の記号処理システムを開発中であり、オブジェクト指向の知識表現を行おうとしている。最終的にはそれを、SPMSのようなアプリケーションシステムのユーザーインターフェイスを実現するために適用したい。これらのシステムの作成を通じて考えるようになった事は、ソフトウェアシステムは、さまざまな専門レベルに対応する概念レベルの階層を重ね、専門知識を各階層内のプロセス、階層間のインターフェイスとして実現しなければならないということである。そのような形態を取らなければ、各専門家がシステムの開発に直接関与することが出来ず、専門知識の早急な集大成が出来ない。我々が直接人間で、相互に各レベルの知識、プログラム、データ、解析法等を交換することの出来る共通の環境ができることが望まれる。

Key words: Statistical package; Portable software; User interface.

2章では、現在使用されている統計パッケージを比較する。特に重要な点については、例に基づいて比較する。3章では、統計パッケージの構成について述べ、その中で特に重要である、データ管理について4章で、ユーザーインターフェイスについて5章で述べる。6章では統計パッケージを公共の物とする可能性について述べる。

2. 現在の統計パッケージ

既に100以上の統計パッケージが開発されている。重要な機能が取込まれてきた過程で、統計パッケージを世代分けしてみる。第1世代はBMDPに代表され、各統計手法を単独のアプリケーションプログラムとして集めたものであり、利用者は統計パッケージとOSとの間を往復しなければならない。第2世代はSPSSに代表され、統合したアプリケーションシステムとなり、前処理と解析を連続したプロセスとして実行できる。しかし、単一の、良く整理された表のみが解析の対象である。第3世代はSASに代表され、強力な表操作機能を前処理として使用でき、複数の表を元にして解析を行うことができる。しかし、解析の道筋は確立していなければならない。第4世代は、Sに代表され、統計手法とデータ操作を一般的な制御の下で可能にしたものである。統計用汎用言語と言っても良く、統計のためのデータタイプがいくつか定義されている。しかし、データ間の関係は記述することができず、統計手法はあっても、統計に関する概念や、対象モデルの表現はできない。第5世代が次世代の統計パッケージである。

2.1 SPMSのシステムとコマンドの概説

SPMSは東京都臨床医学総合研究所において、主記憶96KB、ディスク15MB×4という構成のバッチ用の汎用コンピュータ上でPL/I、FORTRAN及びアセンブラを用いて開発した。幸いなことは、利用者である医者、統計家、システムエンジニアの三者が揃っていたことである。それぞれが自分の主張を行った結果、各部分の独立性の高い、良い構造のシステムとなった。その際開発上の障害となったのは、主記憶が小さいこと、言語に再帰呼び出しの機能がないこと、言語間通信の機能が弱かったことである。そのため、ソフトウェアで統計データに対し仮想記憶機構を作成し、オーバーレイコントロールもプログラム内で行わなければならなかった。これらの障害を克服するための努力がシステムの移植性を悪くした。

下図はSPMSのプログラムの例である。SPMSの基本的データ構造は、RELATIONと呼ばれる項目×ケースのテーブルである。^{*}に続く名前がRELATION名または項目名である。!Nはこのような名称を保持する変数である。これだけの説明があれば、自然語に近い記述法なので、プログラム例の意味は理解できよう。

利用者は、!NのようなCOMMAND VARIABLE（実行時の一次変数）によりコマンド間のデータの受け渡し、実行の制御ができる。大きなデータは項目やRELATIONで受け渡す。

統計家は、利用者に疑似自然語のコマンドを提供するために、統計手法プログラムとの間にインターフェイスプログラムを作成しなければならない。そのとき、基礎システムからは、コマンドを再読み込みのできるトークン列として受け取る。SPMS内のプログラムでは（RELATION名×項目名×ケース番号）を指定することにより、任意のRELATIONの任意の要素を使用でき、RELATION及び項目の生成消滅が自由にできる。すなわち、プロ

グラム上は大きな三次元アドレスが使用できる不揮発性仮想メモリがあるとしてよい。

その他の特長としては、関係演算による複数 RELATION からの RELATION の作成、2つの RELATION を利用した、不定繰り返しの時系列データの表現が挙げられる。

```
REPEAT !N = ^EXPERIM , ^CONTROL ;
  TITLE IS 'CANCER EXPERIMENT DATA : RELATION ' !N ;
  CREATE RELATION !N FROM CARD BY DATA DEFINITION :
    ITEMS ARE ^GOT V(1,4), ^GPT V(6,4), ^WBC V(11,5),
      ^RBC V(17,5), ^CH50 V(23,4,1), ^CANCER-W V(28,4)
    MISSING IS BLANK FOR ALL ;
  IN RELATION !N ;
  MAKE ^LOG(GOT) = LOG10(^GOT + 1.0) ;
  MAKE ^LOG(GPT) = LOG10(^GPT + 1.0) ;
  SHOW DATA OF ^CANCER-W ^CH50 ^WBC ^RBC ^GOT ^LOG(GOT) ^GPT ^LOG(GPT)
    WHERE UNDEFINED SYMBOL = 'MISSING' ;
  PLOT HISTOGRAM OF ^LOG(GOT) ^LOG(GPT) ^WBC ^RBC ^CH50 ;
  APPLY PRINCIPAL COMPONENT ANALYSIS ON ^LOG(GOT) ^LOG(GPT) ^WBC ^RBC ^CH50
    AND SAVE SCORE OF 2 COMPONENTS BY ITEM ^COMP-1 ^COMP-2 ;
  APPLY CLUSTER ANALYSIS OF CASE USING ITEMS ^COMP-1 ^COMP-2 ;
END REPEAT ;
IN RELATION ^EXPERIM ;
MAKE ^CAN-CLAS BY MAPPING FROM ^CANCER-W
  WITH INTERVALS (0.5, 15.0, 60.0) ;          % STRATIFY TO 0,1,2,3
PLOT TYPE 2 GRAPH BETWEEN ^COMP-1 AND ^COMP-2 GROUPED BY ^CAN-CLAS
  WITH PRINTED SYMBOL(0->'N', 1->'A', 2->'B', 3->'C') ;
COMPUTE ELEMENTAL STATISTICS OF ^CH50 ^WBC ^RBC ^GOT ^LOG(GOT) ^GPT
  ^LOG(GPT) ^COMP-1 ^COMP-2
  SELECTED BY ^CAN-CLAS ;
APPLY EQUAL MEAN TEST BETWEEN GROUPE ^CAN-CLAS (1,2,3)
  FOR EACH ITEM ^CH50 ^WBC ^RBC ^LOG(GOT) ^LOG(GPT) ^COMP-1 ^COMP-2
  AND LIST SIGNIFICANT CASE ONLY ;
SPMSによるプログラム例
```

2. 2 現在使用されている統計パッケージの比較

現在使用されている統計パッケージのうち、PRODAS、P-STAT、S、SASとSPMSを比較する。目的は、パッケージの提供する基本的概念を明らかにすることである。有用な統計パッケージであるためには、多くの手法が利用可能でなければならないが、ここでは問題としない。手法を組込むことが容易であれば、あとは単に努力の問題だからである。SとSPMS以外はマニュアル上の比較なので誤解があるかもしれない。

(1) 実行方式

PRODAS、P-STAT、S、SAS：会話型、バッチ型

SPMS：インタープリター（またはCのプリプロセッサ）でバッチ処理

SPMSを除き会話型の処理ができる。SASは効率を上げるため、データステップでコンパイルを行うが、処理速度が早いので、会話の流れを乱すことはない。

(2) 言語の型

S：関数型、統計手法も関数で、結果をそれぞれの型の値として返す。

PRODAS、P-STAT、SAS、SPMS：コマンド型、実行結果は明示されたデータセットに格納される。

S以外では、コマンドレベルでデータ型という意識を持っていない。従って関数合成の形式を取れない。しかし、統計手法は大きな計算であり、多入力多出力なので、その出力を単一のデータ型とする事には少し無理があるように思う。

(3) 判定、繰り返し等の一般的制御

S、SPMS：コマンド言語として可能。

PRODAS、P-STAT、SAS：テーブルを処理する所でのみ可能。

S、SPMSは判定、繰り返し、数値や文字列に対する演算といったプログラム言語での基本機能ともいうべきものを持っている。特にSは各種のデータタイプを持ち、豊富な演算とともに汎用プログラムのスタイルをしている。SPMSの一バージョンをCのプリプロセッサとして作成したのも、基本的なデータと制御は常に必要であり、単純な概念は単純に使用できるべきだとの思想に基づく。

PRODAS、P-STAT、SASでは、それぞれ PPL, program step, data step という名称の、テーブルデータをケースまたはバリアブルについて処理する機能ブロックでのみ一般的制御ができる。

(4) 言語の拡張、又はマクロ機能

S：マクロ展開の中で、一般的な制御、演算、マクロ定義、再帰呼び出しができる。

SAS：コマンド言語の実行制御の弱さをマクロ機能で補っている。すなわち、グローバルマクロバリアブルは常に有効であり、これによりコマンド間のデータの受け渡しを行う。またマクロ展開の中で、Sと同等の機能が使用でき、多数のコマンドを生成することで実質的に繰り返しを行うことができる。

P-STAT：マクロ展開中に文字列の置換が可能である。すなわち引数を持つマクロを書くことができる。

PRODAS：マクロ変数を持つ。

SPMS：無し。

SとSASではマクロ機能が強力なので、メニュープロセッサをマクロで定義できる。

PRODASではメニューというコマンドを持つ。

(5) 手法のパラメータとしての変数値の使用

S：一般の式が使用可能。

SPMS：名称変数を含むスカラー変数が使用可能。

PRODAS、SAS：マクロ変数が使用可能。

P-STAT：使用できない。

(6) データ構造

S:ベクトルと構造で全てのデータを表現する。配列、マトリクス、時系列、カテゴリ、テーブル、プロット、モデル、インデックス、クラスター、距離は特別にシステム内部で使用する構造体タグを与え、システム定義データタイプとする。

PRODA S、P-S T A T、S A S:ケース×変数の表で全てのデータを表現する。計算結果は変数の使用法をシステムで定めた形式の表に格納する。

S P M S:表と、配列が利用できる。

(7) 表操作機能

P R O D A S、S A S:複数の表に対し、一般的なプログラムで操作ができる。特にケースおよび変数の選択が容易である。S O R T、M E R G E を持つ。

P-S T A T:単一の表に対し、一般的な操作ができる。

S P M S:S O R T、P R O J E C T、M E R G E、C O N C A T E N A T E コマンドがある。

S:表は各変数のベクトルを要素とする構造体、あるいは各変数のベクトルの暗黙の集まりとする。直接要素データを扱うコマンドを使用する。ベクトルに対する S O R T、M A T C H 関数を利用する。

(8) 欠損値

P-S T A T:数値、文字列共3種の区別ができる。

S A S:数値は27種、文字列は1種の欠損値が表現できる。

S P M S:数値、文字列とも1種の欠損値を持つ。

P R O D A S、S:数値は欠損値を持つ。文字列は空列となる。

(9) 拡張性

S:計算のプログラムを作成すれば、Sのインターフェイス記述言語により、コマンドを定義できる。Sの内部で使用しているプログラムは全て利用できる。但し、コマンドのシンタクスは位置又はキーワードによる引数の記述に制限される。データがそのタイプ、構造に関して自己記述的になっているので、新しいデータタイプを導入して、Sそのものを拡張することが可能と思われる。

S P M S:語い解析ルーチンとのインターフェイスを用いて、自分でコマンドの解析を行う。従って全く自由なシンタクスが可能である。内部利用のプログラムは全て利用できる。

S A S:システムで定めた引数だけが自動的にユーザプログラムに渡される。その他はコマンド文字列そのものが渡される。

P-S T A T:可能(The P-STAT Technical Guide)

(10) ヘルプ機能

S:関数の他にデータセットについての説明も行う。

P-S T A T、S A S:コマンド、キーワードの説明が得られる。

(11) ジャーナル機能

S、SAS：入力、出力とも選択的に保存できる。

PRODAS：入力は再利用可能である。

2.3 コマンドの例による比較

統計パッケージ間の大きな違いはデータの使用法に見られる。具体例により各パッケージを比較する。

第一の例は、二つの表を併合し、データ項目の値により出力のタイトルを設定する。

```
inx <- match(SHIPMNTS$ID, PRODUCTS$ID);
REPORT <- cstr(REPTIME =SHIPMNTS$REPTIME ,ID=SHIPMNT$ID,
               MODEL=PRODUCTS$MODEL[inx],INVENTORY=PRODUCTS$INVENTORY[inx]);
if(REPORT$REPTIME[1] == "MONTH") PERIOD <- "CURRENT MONTH" else
  PERIOD <- "CURRENT YEAR";
print encode("REPORT FOR", PERIOD);
print REPORT;
```

Sによるプログラム例

```
PROJECT ^SHIPMNTS ON ^TEMP1 WITH ITEMS ^REPTIME ^ID ;
PROJECT ^PRODUCTS ON ^TEMP2 WITH ITEMS ^MODEL ^ID ^INVENTORY ;
MERGE ^TEMP1 AND ^TEMP2 TO ^REPORT BY KEY ^ID ;
ASSIGN IS AS FOLLOWS:
  IF DATA(^REPORT , ^REPTIME ,1) = "MONTH"
  THEN "CURRENT MONTH" ELSE "CURRENT YEAR" ;
TITLE IS "REPORT FOR " IS ;
IN RELATION ^REPORT ;
  SHOW DATA OF ^ID ^MODEL ^ INVENTORY ;
```

SPMSによるプログラム例

```
DATA REPORT ;
  MERGE SHIPMNTS(KEEP=REPTIME ID) PRODUCTS(KEEP=MODEL ID INVENTORY);
  BY ID ;
RUN ;
%MACRO TIME ;
  %GLOBAL PERIOD ;
  DATA _NULL_ ;
  SET REPORT(OBS=1);
  CALL SYMPUT('T',REPTIME);
  RUN ;
  %IF &T=MONTH %THEN %LET PERIOD=CURRENT MONTH ;
  %ELSE %LET PERIOD=CURRENT YEAR ;
%MEND TIME ;
%TIME
PROC PRINT DATA=REPORT;
  TITLE "REPROT FOR &PERIOD";
RUN ;
```

SASによるプログラム例

表の併合の意味はシステムにより違いがある。表を一括して扱う機能では、SASが最も強力かつ高速である。Sの場合、表がデータタイプではないので直接利用できるコマンドがない。一方SASではコマンドからデータ値を直接利用することが著しく困難である。PSTAT、PRODASではコマンドレベルで利用できる変数値を保持する機構がないのでこの例の記述はできないように思われる。

次は不定繰り返しのデータから計算を行う例である。個人と測定値の二つの表からなり、個人表に平均値を付加する。二つの表はSPMSではLINKコマンドにより直接、SとSASでは個人IDにより暗黙に結合されているとする。

```
INDX <- code(MEASURE$ID, PERSON$ID) ;
PERSON <- mstr(PERSON,WEIGHT=tapply(MEASURE$WEIGHT,INDX,"mean"))$Data) ;
```

Sによるプログラム例

```
IN RELATION ^PERSON LINKED TO ^MEASURE ;
MAKE IN BASIC RELATION MEAN ITEM ^WEIGHT OF ^WEIGHT ;
```

SPMSによるプログラム例

```
PROC MEANS DATA=MEASURE ;
  BY ID ;
  VAR  WEIGHT ;
  OUTPUT OUT=TABL MEAN=WEIGHT ;
DATA TMP ;
  MERGE PERSON TABL ;
  BY ID ;
PROC DATASETS ;
  CHANGE TMP=PERSON ;
```

SASによるプログラム例

これらの例から、データの表現法が直接コマンドに影響しており、複数の表が関係することの概念的把握はなされておらず、操作的にしか処理できないことが分かる。次世代の統計パッケージでは、ネットワーク構造のデータの上に、任意の統計モデルあるいはビューを表現できなければならない。

3. 専門家集合としての統計パッケージ

統計パッケージの構造を考えて見ると、各種の専門領域が関連していることがわかる。利用者、統計家、数値計算家、言語専門家、OS専門家、データベース専門家、グラフィックス専門家等である。同一の物でも専門が異なれば異なる意味を持つ。同じような概念も、専門が異なれば、多数の概念に分類される。我々は複数の専門に精通することは少ないので、このような複雑なシステムを作成するためには、明白で安定したインターフェイスが必要である。これは、各専門での概念について、用語、意味、用法が明確かつ認められていなければならないことを意味する。FORTRAN等の言語で記述したプログラムは、記述法が言語の規約に縛られるため、この条件を満たさないことが多い。従って、専

門用語で記述したプログラムを計算機言語に翻訳する機構が必要であり、逆に、この機構によって専門用語のオペレーショナルな意味が表現される。

利用者の立場から見ると、統計パッケージが持つ、データの性質を記述する言葉が少ない事に気づく。これは、統計パッケージが多くの手法を提供することを第一目的とし、計算プログラムが受け入れる形にデータを整えること、計算プログラムが理解し易い表現で指令することを利用者に強制して来たからである。

Sを除くと、ほとんどの場合、フラットテーブル上に配置されたスカラーデータのみを対象としており、表形式以外のデータを扱おうとすると、表現と意味との解離を招く。一見高度に階層化されているかに見える統計パッケージも、実は概念的には浅いシステムでしかない。高度な統計手法というマクロなオペレーションを実行し、手法を適用した結果が特に重要な概念であったために、誤解されているにすぎず、概念レベルは浅いのである。従って、利用者がデータの持つ残りの概念を記憶していなければならない。Sにおいてもデータ間の関係の管理は利用者にゆだねられている。

次世代の統計パッケージでは、それを構成する各レベルにおいて、対応する各専門家の持つ概念で記述ができなければならない。利用者は統計的概念を用いて記述し、統計家は数学的概念を用いて新しい手法を記述する。各レベルではそれぞれ、提供する概念について、それが「何」で、それを実現するために必要なことを下位レベルの概念から「どのように」構成するか、その方法を提供する。通常はサブルーチンライブラリとして提供されるものであるが、概念は文脈に依存し、プロセヂアとデータの組み合わせで表現される。従って、提供の形態は、単なるサブルーチンライブラリでは不十分である。各レベルごとに言語を管理し、利用者の要求に応じて言語の定義、レベル間の翻訳を行い、利用者に対し意味の説明を行う専門家プログラムに内蔵された形態で提供することが必要である。

また概念に必然的に付随する属性については、常に利用可能であるようにする。たとえば、統計レベルで interval (Stevens's Measurement System) と概念規定された変数は、数学レベルでは実数の集合であり、平均値という属性を持つ。従って利用者は、いつでもこの変数の平均値を利用できる。値そのものは、数値計算レベルに適するように、一次元配列として記述され、物理的には、アクセス効率を考慮して配置されるであろう。データは論理的には下図のように、データ自体が自分に関する記述を各専門レベルごとに保持し、上位概念から限定される属性、自分の属性から導かれる値等を持つ。図では複数のレベルを同じ形式で記述したが、必然ではない。また、概念は単にデータ表現だけで記述できるものではない。その概念に付随するプロセヂアが規定され、そのプロセヂアを通してのみ意味を持つことが重要である。すなわち、統計データは抽象データタイプ、あるいはオブジェクトの一種として考えなければならない。

統計の目的は対象を理解することにある。モデルとの適合性、検定等は、理論的理解と言える。作図、作表等で観測した事実を基に、対象の性質が付加され、あるいは異なるモデルに変換される(統計手法を適用することは、対象に新しいビューを付加することである)。こうして形成された対象そのものが解析結果であり、必要な情報は、対象に問いかけることにより、直接に得られる。統計パッケージを構成する各専門家は、この対象の記述を媒介にして起動される。

```

(OBJECT :
  (NAME:   health)
  (CLASS:  set_of_measurement)
  (NO_OF_OBS: 1000)
  (ITEM:
    (NAME:   height)
    (MEASUREMENT_CLASS: interval)
    (CLASS:  indexed_set)
    (NO_OF_MEMBERS: 1000)
    (MEMBER_CLASS: lenght)
    (UNIT:   feet)
    (VALUE:
      (MEAN:   5.36)
      (CLASS:  set_of_real)
      (EXISTANCE:
        (LOCATION: data/health/height)
        (DESCRIPTION: vector_of_double)
      )
    )
  )
  (ITEM: ... )
  ...
)

```

内部データの記述例

4. データの表現と管理

一般的に、データベースにはデータの蓄積更新、一括読みだし、検索の機能が要求される。それぞれを特徴とするシステムには銀行のトランザクション処理システム、統計パッケージ、文献検索システムなどがある。データベースをデータとそれを表現するスキーマで捕らえるのが一般的である。トランザクション処理システムではスキーマは安定し、データが急速に更新される。統計パッケージでは、データは変化せず、スキーマが変化する。文献検索システムでは、データ、スキーマとも変化しないが、膨大なデータ量に対し早いレスポンスタイムが要求される。このように、これら3種の機能は全く異なる特性を持つので、単一のデータ管理法で効率良く実現するのは不可能と思われる。統計パッケージのデータ管理は、データ値だけでなく、スキーマすなわち概念、及びその形成過程を表すデータの管理を行わなければならない。

ここでは、物理レベルにおけるデータの問題を考える。どのように論理的なデータ設計をしても、全てのデータは外部メモリに置かれるという状況は変わらないと予想される。またCPUと外部記憶の速度差も変わらないであろう。システムの効率に最も影響があるのはデータ管理である。計算専門家プログラムはデータ管理専門家を介して得た主記憶上の値を用いる事とすると(仮想記憶が必要)、データ管理専門家のもっぱら効率を考えて物理データの管理を行えば良い。

統計パッケージにおいては、ケースレコードの蓄積更新が中心となる一次入力データの管理は行うべきでない。もっと一般に、統計パッケージではデータの更新そのものを行う

べきではない。これは、更新という処理が、理由、根拠を明示せずに、結果の値だけを与えるからである。データが一旦統計パッケージ中に入ったら、新しいデータの生成以外は許してはならない。データベースには色々な見方からのデータが蓄えられており、その多くは加工して得られた二次データである。二次データは全てそれがどのようにして得られたか、その履歴を持っていて、正当性を表出することができなければならない。単純な更新は二次データの正当性を失わせる。

データ管理専門家の役割は主に次の4つである。1. ファイルと主記憶間のデータのマッピング、2. 効率的物理データの配置、3. 不用データの処置、4. 名称と実体との対応の管理。

更新を許さないことにより、データの物理的コピーを許す事ができる。また集合データは全体を一時に利用する事が多い。単に既存のデータの組み合わせを行う事により新しいデータビューを作るときは、データを共有することができる。これら事実を用いて、データ管理専門家は効率良い物理データ表現をすることができる。

データ管理専門家のもう一つの重要な任務は、不用データの削除である。現在使用中の名前でなく、どの二次データの支持データでもないデータは削除して良い。また、長期間使用されていない二次データは、必要なら履歴データから復元すれば良いので削除できる。

名称と実体との対応はあいまいにされがちである。名称そのもの、現在ある名称で呼ばれているもの、ある実体につけられた名称など、色々な名称は用いられる。しかしこのあいまいさのなかに利用者の概念が潜んでいる。データベース管理者は、名前が多重に使用されることを許し、実体の表す概念、生成された時刻で同定を行い、利用者にはあいまいだが理解しやすい名前による表現と、完全に個別のデータを指定できる内部的表現の両方を提供する。

データとして表現される事が増大するので、現在以上に二次記憶装置へのアクセスが増加する。不定形、可変長のデータへのランダムなアクセスと、大規模配列データへのアクセスである。これらをどのように効率的に行うのかがシステムの性能を決める。これから主流となる個人利用の仮想記憶を持つワークステーションを想定すると、I/Oの回数、I/O待ちの時間を減少させることが重要である。I/O及び仮想記憶を管理しているのはOSである。統計データファイルを入力せずに直接にメモリセグメントとして利用できる機構、及び仮想記憶上でのシーケンシャルアクセスを検出して次ページをプリフェッチする機構、をOSが持つことを期待する。

5. 統計家の弟子

理想的な統計家の弟子とは、計算機システム内に創造された利用者の分身である。それは、あらかじめシステム内の各専門家に關する知識を持ち、利用者と各専門家との間の通訳を行う。利用者が与えた知識により、個人的な環境を作成する。それはヘルプ機能、ヒストリー機能、マクロ機能、デフォルト機能等のユーザインターフェイスを統合化し、利用者の概念形成を支援する機能を付加したものと考えることができる。統計家の弟子の下で、データ及び手法は利用者の言葉と結合した形の存在となる。現在の統計パッケージの会話型とは、コマンドごとに実行結果やエラーメッセージが端末に表示されるという意味にすぎない。真に会話的とは、会話により理解に達すると、お互いに相手の思考に關す

る知識を獲得することであろう。そうなれば次のようなことが可能となる。

会話の中で理解できないことや不明確なことがあれば、自動的にそれを解決するための会話が行われ、その結果として、新しい知識が獲得される。その過程において、利用者の知識とパッケージ内専門家の知識が照合される。すなわちヘルプ機能は単にパッケージのキーワードの説明を得るだけではなくなる。

利用者の要求により新しいデータを生成するとき、統計家の弟子は、要求を統計専門家の形式に変換するとともに、データに命令と支持集合を付加する。これにより、データが一次データから導かれて来た経路と、そのデータの影響するデータを示すことができる。さらに支持集合の一部を変数化することにより、マクロ機能を定義することができ、概念形成の補助となる。

統計家の弟子は、利用者の概念、データの性質、手法の知識、文脈等を知った上で利用者の要求を統計専門家に伝えるので、セマンティックギャップがほとんどなくなる。また統計家の弟子の獲得した概念は、それ自体可搬であり、流通する知識となる。

6. 共有財産としての統計パッケージ

統計パッケージには多くの専門家の知識が集積されている。知識は共有されてこそ価値があるので、我々はデータのみならず、統計パッケージを構成する各部分を共有し、自由に組み合わせて利用できるようにすべきである。可搬性の高い、階層の多いシステムは多くのハードウェア資源を使用し、一般には実行の効率が悪くなる。しかし統計パッケージのように、マクロな機能が標準化されていれば、効率はあまり低下しない。近年使用されるようになった個人用ワークステーションは、計算速度、主記憶、補助記憶とも充分な能力を持っている。利用形態が個人的になればなるほど知識の共有化、すなわち、データ、プログラム等を交換することが重要になる。これは一種の通信であり、統計パッケージに必要なハードウェア、OS等に依存する事項の記述を標準化する必要がある。Sが個人間で知識を共有する媒体となれるかどうかを調べてみる。

SはUNIX系システムで利用できる。ハードウェア、OS及びコンパイラの違いは、約60のパラメータで記述され、UNIXの範囲内では移植性に富んでいる。特に、システムの組込法が自動化されているので、利用者が自分でシステムの生成を行うことができる。また、このシステム生成の方法は、利用者定義の機能を組込む方法と同じであるので、利用者による機能追加も容易である。

Sでは190の手法、14のグラフィック装置が使用できる。内部は、計算、データ管理、グラフィックス、言語処理に大別され、これらを記述するために、1400のプログラム、35の制御エリアと950の定数（とローカルなシンタックス変換器）が定義されている。内部の機能を利用、作成するには複雑過ぎるほどの規模である。

以上から、Sは内部専門家プログラムによる援助が得られるようになれば、知識を共有する媒体となる可能性がある。

7. まとめ

現在の統計パッケージの機能、構造を調べた結果、手法の集まり以上を目指すには、含まれている概念が低レベルであり、その上概念形成の能力を欠いていることがわかった。

統計パッケージを更に有用なものにしていくためには、システムの階層性を専門家と対応させて明確にし、各階層におけるデータ表現、階層間のインターフェイスを確立し、独立した個人の生産した知識が、計算機上で結合できる、共通の環境を確立しなければならないことがわかった。この環境がハードウェアやOSに依存しないためには、ワークステーションと呼ばれるクラスのハードウェアの能力が必要である。しかし、それらの価格の低下は著しく、多くの研究者、利用者が使用できる時代になりつつある。次世代統計パッケージでは、利用者は統計家の弟子と呼ぶ統合されたユーザーインターフェイスにより、システム内の各専門家と対話ができ、その結果、概念形成、プログラム作成等において、個人的環境での援助を受けられる。また、データ、プログラム、概念等を共通の自己記述的プロトコルに変換して交換することができる。すなわち、次世代統計パッケージは社会的規模で知識を蓄積する環境を提供するものである。

本論文をまとめるにあたって議論をしていただいた富田教授に感謝します。

8. 参考文献

- Afifi, A.A. and Clark, V. (1984). Computer-aided multivariate analysis., Lifetime Learning Publications.
- Becker, R.A. and Chambers, J.M. (1984). S An Interactive Environment for Data Analysis and Graphics, WADSWORTH.
- Becker, R.A. and Chambers, J.M. (1985). Extending the S System, WADSWORTH.
- Bobrow, D.G. and Stefik, M. (1983). The LOOPS Manual, Memo KB-VLSI-81-13, Xerox Corporation
- Date, C.J. (1981). An Introduction to Database Systems, addison-wesley
- Doyle, J. (1979). A Truth Maintenance System, Artificial Intelligence., Vol.12, No. 3, 231-272
- Francis, I. (1981). THE STATE OF EXISTING STATISTICAL SOFTWARE, INTERNATIONAL ROUNDTABLE CONGRESS, THE 50-TH ANNIVERSARY OF THE JAPAN STATISTICAL SOCIETY, A13-A27
- Kariya, J. and Tango, T. (1979). The design and implimentation of SPMS (Statistical Package for Medical Science), Proc. 12th Annual Symposium on the Interface, 476-480
- P-STAT (1985). P-STAT 8 User's Manual, Third Ed., P-STAT Inc.
- PRODAS (1985). PRODAS Referece Manual, Version 3.0, Conceptual Software Inc.
- PRODAS (1985). PRODAS User's Guide 3.0, Conceptual Software Inc.
- Rohlf, F.J. (1984). USE OF PERSONAL COMPUTERS IN STATISTICS AND BIOLOGY., Proc. The XIth International Biometric Conference., 139-147
- SAS (1985). SAS User's Guide: Basics, Version 5 Ed., SAS Institute Inc.
- SAS (1985). SAS User's Guide: Statistics, Version 5 Ed., SAS Institute Inc.
- SAS (1981). SAS PROGRAMMER'S GUIDE, 1981 ED., SAS Institute Inc.
- Schueier, B. (1977). UPDATE RECONSIDERED., Architecture and Models in Data Base Management System, North Holland, 149-164
- Seels, L. (1979). Reasoning Modeld As A Society of Communicating Experts., AI-TR-542, MIT
- Stonebraker, M. (1981). Operating System Support for Database Management, CACM, 412-418
- Tango, T. (1980). SPMS(Statistical Package for Medical Science) User's Guide.
- [Received November 1986]