

統計解析システムSASの紹介

—SAS言語を中心として—

新村秀一 住商コンピューターサービス株式会社

1. SAS紹介

SAS (Statistical Analysis System) は、1966年にノースカロライナ州立大学 (NC SU) で A. J. Barr, J. H. Goodnight らを中心として開発されました。現在は、これらの開発者が中心となり設立した、SAS Institute Inc. により開発・販売されています。

SAS と他のパッケージを一言で比較するならば、他の統計パッケージは『パッケージの域をでていない』ということです。これに対して SAS は、『統計処理もできる高級言語』または『データ解析のための一大システム』と考えないと正しく理解できません。すなわち、①入出力媒体に制限がないこと (COM やグラフィック・ディスプレイへの出力可)、②大規模データファイル (データ件数は無制限、変数数は 4000 迄) または、IMS や ADABAS 等のデータベースの処理が可能なこと、③PL/1 言語に似た SAS 言語でユーザーが自由にコーディングできること、④SAS 言語のもとで各種ユーティリティ (ソート・マージ・アップデイト等) の利用やシステムオプション、関数等が利用できること、⑤TSS の導入が容易なこと、等の特徴を有します。

このことは、SAS ユーザーが単に大学や研究機関にとどまらず、金融業等の一般企業にもインストールされていることと無関係ではないと考えられます。また、統計パッケージとしては SAS のみが 1976 年から連続して Datapro Research 社の最優秀製品リストに選ばれています。現在のユーザー数は 1126 ケ所 (各四半期毎に 10% から 25%

の増加率) であります。

2. SAS の概要

使用可能機種: IBM 360/370 とその互換機

使用 OS: OS と OS/V S

使用メモリ: 240KB 以上 (ユーザーエリアは 150KB 以上)

総ステップ数および使用言語: 12 万ステップ以上。50% アセンブラー (re-entrant), 48% PL/1 (structured), 2% Fortran (old program)

主機能: 統計解析およびデータ処理。

データマネジメント。レポートライティング。検索。

プログラム・ライブラリ: 3330 disk で 400 トラック (5.2 メガ)

モジュール数: 300 以上のロードモジュール

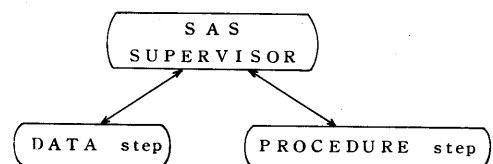
その他: TSS 可能。カラーディスプレイやプロッター上へのグラフ出力。データベースへのアクセス。

3. SAS 言語

3-1 SAS 言語の構成

SAS 言語は、図 1 に示すように、

図 1



① SAS データセット (システムの標準ファイル) を生成・管理する DATA

step と

② PROC step で定義されるデータ解析用ライブラリー

から構成され、SAS スーパーバイザーの管理のもとで動く。

全ての SAS Job は、DATA step もしくは PROC step のいずれかで始まるステップ群の一連のつらなりからできており、複数 job を連続して実行することも解析結果を SAS データセットまたは外部ファイルに出力し再度計算することも容易である。

3-2 SAS 言語の文法

SAS 言語の特徴としては、その文法が非常に単純明快なことである。加えて PL/I 形式のフリーフォーマット記述であること、default 値が設定してあることがユーザーにとって使い勝手の良いものになっている。

多くのパッケージがたとえその処理内容が良いものであったとしても、ユーザーから拒絶されるのは、パラメーターやコマンドの体系を人間工学の立場から一つの分かり易い文法を作るのだという認識に乏しいからではなかろうか。たとえばカラムを指定するという方法が良く採用されている。しかし、この方法では意外とカラムずれのため実行が不可能になったり、そのパッケージにたとえ熟練していたとしてもカラム確認のため必ずマニュアルを参照する必要にせまられる。

1) DATA step

SAS データセットの生成・変更・管理を行なう。

1-1) 外部データを入力する場合

```
DATA data-set-name ;
INFILE dd-name ;
INPUT data-description ;
optional program statements ;
CARDS ;
```

(input data lines)

外部データを、カード、TSS ターミナル、MT、ディスク等から入力し、SAS データセット (システムの標準ファイル) を作成します。

INFILE ステートメントで定義した dd 名の入力媒体から、INPUT ステートメントで定義したデータ記述に従ってデータを 1 件ずつ入力し、オプション・プログラムでデータの加工を行ない、DATA ステートメントで定義したデータセット名をもつ SAS データセットを出力します。

1-2) 既存の SAS データセットを入力とする場合

```
DATA data-set-name ;
{ SET data-set-names
  or
  MERGE data-set-names
  or
  UPDATE data-set-names ;
  BY variable-names ;
  optional program statements ;
```

既存の SAS データセットから、新しい SAS データセットの作成を行ないます。

SET ステートメントにより、1 個または複数の SAS データセットが入力され、1 個の新しい SAS データセットが作られます。MERGE と UPDATE ステートメントにより、入力した複数の SAS データセットのマージとアップデイトを行ない、その結果を新しい SAS データセットとして出力します。

1-3) データの外部媒体への出力

```
DATA _NULL_ ;
SET data-set-name ;
( or another source of input )
data
```

```
FILE dd-name ;
optional program statements ;
PUT variables and formats ;
```

SETステートメントで定義されたSASデータセットを入力し、FILEステートメントで定義されたdd名をもつ出力媒体へ、オブショナルプログラムでデータを加工後、PUTステートメントで与えた書式に従い出力する。

2) PROC step

PROC stepは、PROCステートメントと、それに付随する情報ステートメントから構成される。情報ステートメントは、PROCステートメント中のオプションやパラメータでは記述しにくい情報又は機能の詳細記述を行なう。どのモジュールにも共通して用いられる変数記述のステートメントと、特定モジュールで意味をもつ特殊ステートメントがある。

```
PROC module-name option
parameter data-set-reference ;
VAR list-of-variables ;
BY list-of-variables ;
ID variable ;
MODEL model-description ;
TABLES table-request ;
PLOT plot-request ;
OUTPUT data-set-name ;
etc.
```

情報ステートメント

PROCステートメントはPROCで始まり、その後には必ずFACTOR(因子分析)やNLIN(非線型回帰)等のモジュール名を伴う。オプション、パラメータ、入力データセットの指定順は任意である。因子分析を例にとれば、18個のオプションとパラメータがあり、METHOD=ALPHAで6種の手法からアルフ

ァ解析を指定しPLOTと書くことにより因子得点のプロットを行なう。更に多くの場合、システムのdefault値が設定されている。例えば、DATA=data-set-nameとして処理対象のデータセットを指定するが、省いた場合には最新のデータセットの処理を行なう。

3-3 簡単なSAS jobの例

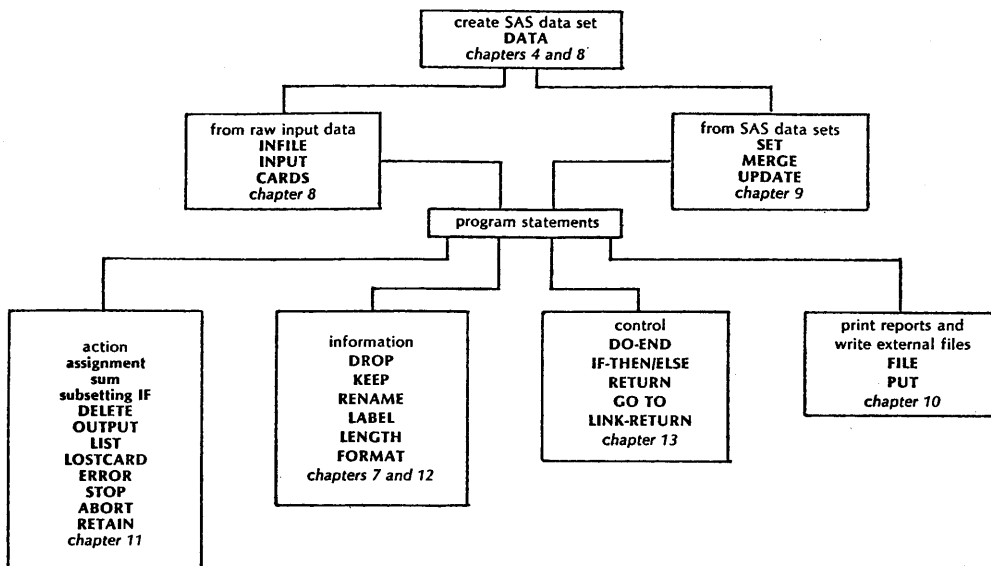
```
DATA ss; INFILE mt;
INPUT (x1-x100) (100
* 5.2);
PROC GLM;
MODEL x1=x2-x100;
PROC FACTOR OUT=f1
METHOD=prin
ROTATE=varimax;
```

dd名がmtと定義された入力媒体から、x1からx100迄の100変数をF 5.2のタイプで入力し、SASデータセットssを作成する。x1を目的変数、x2からx100を説明変数として重回帰分析を行なう。次にx1からx100迄の100変数で因子分析を行なう。

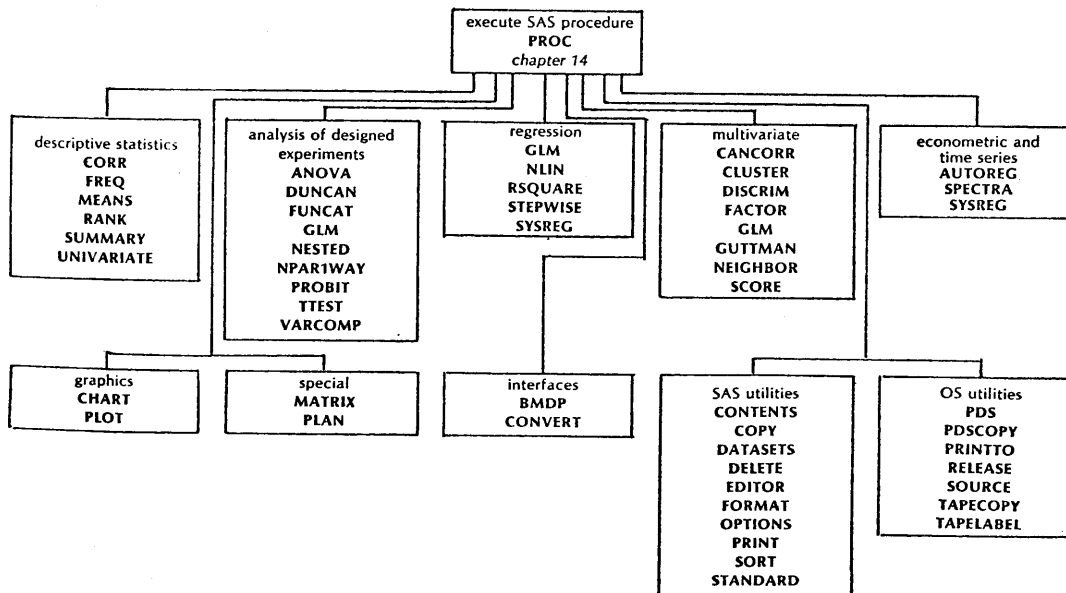
4. SAS言語の体系

図2にSAS言語の体系を示す。この図はSAS USER'S GUIDE(1979版)を主体としている。この他にユーザー開発プログラムやテスト使用中のものとして次のものがある。DATACHK, EXPLODE, FORMS, HARVEY, IPFPHC, KSLTEST, LAV, OPSCAN, PCTL, PREDICT, RSP, SPLOT, TPLSAS, TSCSREG, ARIMA, SIMLIN, X11, MORTGAGE, MODEL, SYSNLIN, SIMNLIN, PRINCOMP. 医学用としては、PHGLM(Cox proportional

DATA step



PROC step



hazards linear model); COXREG (Cox's life-table regression model), LOGIST(logistic multiple regression model), SURVTEST(tests for differences between one or more survival curves), SEQCENS(computes progressively censored tests of multiple regression hypotheses for survivorship data), PREDICT(multivariate logistic model)。

SASは毎年1月に開催されるユーザー会において、当該年度以降の開発計画を決定する。このためSASは年々強かつ巨大なシステムに生長しつづけている。

5. SASのデータと入出力

SASの入出力の特徴は、入出力媒体に制限のないこと、大規模データファイルの処理が可能なこと、入出力命令の多様性にある。

具体的に言えば、数値データの絶対値の大きさは 10^{-73} から 10^{73} 、文字データは200文字まで扱える。SASデータセットの大きさは、4000変数まで、データ数に制限はない。

表1に入出力命令の一覧表を示す。例えばw.またはw.dと記載のあるものは、32. または32.10と書くことにより、Fortran言語でのF 32.0またはF 32.10を意味する。この機能は入出力兼用であり、32文字以内右づめの規則に従う。BESTw.はBEST 10.と書くことにより出力で10文字のエリアがとられ、実データの内容により整数、実数、Eタイプ等のフォーマットをSASが選択するので結果のオーバーフローがさけられる。文字の場合も同様に\$ 200.と指定すれば200文字を1変数で扱える。更に、日付や日時表示等の編集を行なう命令も準備されている。

次に入出力命令に用いられるポインタを説明する。

- ① 現ライン上の指示されたカラムにポインタを動かす。

```
INPUT  cn  @cn  name $10.;
```

変数cnの値が15なら、15カラム目から文字項目nameを10文字分読み込む。

- + 指定されたカラム数だけポインタを動かす。

```
INPUT  @23  length 4. +
5  width ;
```

ポインタは23カラムに移動し、そこから4カラムを数字項目lengthとして入力する。次に32カラムから項目widthを入力する。

- # 指定されたラインの1カラム目にポインタを動かす。

```
INPUT  name $ 1-10  #2
age 3-4  #4 ;
```

1ライン目の1カラムから10カラム迄を文字項目nameとして入力し、2ラインの3カラムから4カラム迄を数字項目ageとして入力する。

この例では、1オブザベーションが4レコードから構成されているので#4と指定することにより3と4レコードを空読みする。

- / 次のラインの始めにポインタを進める。

```
INPUT  a / b ;
```

1ライン目から項目aを、2ライン目から項目bを読む。

- ② 入力ステートメントの最後に用いられた場合、現データラインで停止し、同一DATA step内の次の入力ステートメントの実行に継続される。

```
DATA  schedule ;
```

```
INPUT  type $ 1. ② ;
```

```
IF  type = 'c'▼
```

```
THEN INPUT  class
```

```
instruct time ;
```

```
IF  type = 's'▼
```

表 1

Numeric formats

FORMAT	DESCRIPTION	Input or Output?	Width Range	Decimal Range	Default Width	Alignment
w. w.d	standard numeric	both	1-32	no limit on input; for output, $d < w$	1	right
BESTw.	SAS chooses best notation	output	1-32		12	right
BZw.d	blanks are zeros	input	1-20	0-10	1	
COMMAw.d	commas in numbers	output	2-32	0 or 2	6	right
DOLLARw.d	dollar sign, commas	output	2-32	0 or 2	6	right
Ew.	scientific notation	both	7-32		12	right
FRACTw.	fractions	output	4-20		10	right
HEXw.	numeric hexadecimal	both	1-16		8	left
IBw.d	integer binary	both	1-8		4	left
PDw.d	packed decimal	both	1-16	0-10; for output, $d < w$	1	left
PIBw.d	positive integer binary	both	1-8	0-10; for output, $d < w$		left
PKw.	unsigned packed decimal	input	1-16		1	left
RBw.d	real binary (floating point)	both	2-8	1-10; for output, $d < w$	4	left
ROMANw.	Roman numerals	output	2-32		6	left
SSNw.	social security numbers	output	11		11	
WORDSw.	numbers in words	output	5-200		10	left
Zw.d	print leading zeros	output	1-20		1	right
ZDw.d	zoned decimal	both	1-32	0-10; for output, $d < w$	1	left
ZDBw.	zoned decimal with blanks legal	both	1-32		1	left

Character formats

FORMAT	DESCRIPTION	Input or Output?	Width Range	Default Width	Alignment
\$w.	standard character	both	1-200	1 or length of variable	left
\$CHARw.	characters with blanks	both	1-200	1 or length of variable	left
\$HEXw.	character hexadecimal	both	1-200	1 or length of variable	left
\$VARYINGw.	varying-length values	both	1-200	must give length value or variable	left

Date, time, and date-time formats

FORMAT	DESCRIPTION	Input or Output?	Width Range	Default Width	Alignment
DATEw.	dates of form ddmmyy	both	7-32	7	right
DATETIMEw.	date-time values	both	7-32	18	right
HHMMw.d	hour and minutes	output	2-20	5	right
HOURw.d	hour	output	2-20	2	right
MMSSw.d	minutes and seconds	output	2-20	5	right
PDTIMEw.	packed-decimal time from RMF records	input	4	4	
SMFSTAMPw.	time-date field from SMF records	input	8	8	
TIMEw.d	time values	both	5-32	8	right
TODw.	time-of-day	output	2-20	8	right
WEEKDATEw.	date values	output	16-37	29	right
WORDDATEw.	date values	output	16-32	18	right
YYMMDDw.	date values	both	6-32	8	right

THEN INPUT name id ;
最初のINPUTステートメントで1カラム目から type の値を入力しポインターは2カラム目で停止する。次の2つの1Fステートメントで type の値が参照されcかsかにより別々の項目を入力する。

②② INPUTステートメントの最後に指定すると、現ラインを次のINPUTステートメントの実行のために保持する。
INPUT a b ②② ;
同一ラインから a, b, a, b…の順で複数個の観測値を入力する。

以上入力を例としてポインターの説明を行った。SASの入力はINFILとINPUTの両ステートメントで記述される。これと対称的に出力はFILEとPUTで記述され、データ記述方法やポインターの使用法は同じである。しかし、レポートライティングの便宜のため数ラインをholdして井と②ポインターを用いてランダムに出力イメージを作成後出力する機能もある。

6. 医学領域への接点をもとめて

SASは特定領域に限定されない汎用パッケージであるが、医学領域への適用という観点から考察する。

手法の立場からみれば、一般的な多変量解析やノンパラメトリック検定手法の利用は当然のことながら、数量化I類に相当するFUNCT, プロビット解析のためのPROBIT, 強力な非線型回帰推定のNLIN, 脳波解析やトレンド解析に用いられるSPECTRAやTSCSREG, PHGLMやCOXREG等のMayo ClinicやDuke Medical Centerのユーザー開発プログラム, システム動特性の解析のためのMODELやSYSNLINやSIMNLIN等がある。

医学データベースの立場からみれば、文字

項目が200文字迄許されることの意義は大きい。データ収集はバッチ方式をとらざるをえないが、構築済みのデータベースに対しては強力な処理手段が利用できる。

システムの立場から見れば、各種ユーティリティやデータ保護等の諸機能が簡単に利用できること、専用のレポートライティング機能に匹敵する程の機能を有すること等から、省専門家、省時間、省プログラムがはかれる。

また、ユーザーがSASプログラミングステートメントやMATRIXプロセジャーを用いて自由にプログラミングができるシステムの拡張性も重要な特徴であろう。

7. お わ り に

SASは統計パッケージ利用者に限らず、広くソフト開発に従事する者にとって大きな魅力である。ますます巨大システムに生長しつづけるアメリカ方式に対して、我々は代替案すらもちあわせていないように思われる。

本稿の一部は、当社パンフレット“SASのご紹介”から引用しました。

文 献

- 1) A.J. Barr, et al., A user's guide to SAS 79. (SAS Institute Inc., North Carolina, 1979).
- 2) 川村 昇, 新村秀一, 開発成允: 医学データ解析のための統計プログラムパッケージ, 医学のあゆみ109(11), pp620-626, 1979.
- 3) 新村秀一: 直交表を用いた重回帰分析の変数選択, 1980年度日本OR学会, pp110-111, 1979.