SAS による統計分析入門

東北大学大学院農学研究科　八巻邦次

概要

SAS による統計分析を主にデータ入力、基本的統計手法を学ぶ方法から、多変量の分散分析で複雑な現象の実験結果を解析する SAS プログラムの作り方とそのアウトプットをどう解釈するかを解説した。

1) SAS とは？

東北大学農学部の3年生の学生に動物実験のデータを自ら取り、情報シナジーセンター大規模科学計算システムのアプリケーションである SAS で統計分析して実験結果を発表することを教えてきた。その経験から、SAS による統計分析の入門テキストをここに記してみた。

その SAS (Statistical Analysis System)とは North Carolina State University で開発が始まった統計分析プログラムパッケージで現在では米国 SAS 社が全世界に配布しており、統計パッケージとしては質・量ともに世界で先進的なものといえる。

現在では、単に統計パッケージとしてでなく、さらに発展し民間企業、政府機関、研究機関や大学で、利用される経済予測や製品コントロール、臨床試験やデータベースマーケティング、健康調査、顧客指向調査や株式市場のトレンドに至るまでの情報を評価することができるまでになっている。

SAS の大規模科学計算システムへの導入プロダクトとその利用方法については、小野敏[1)]「ライブラリー・アプリケーションソフトウエアーの紹介」に詳しく書かれているのでそれを参照されたい。今回は、この膨大なアプリケーションプログラムの中から、BaseSAS (SAS の入出力や基礎的な統計計算のパッケージ)、SAS/STAT (高度な統計計算のプログラムパッケージ)の二つを使って、多変量の分散分析を行って、複雑な実験結果の統計分析を行うことを試みた。

2) 大規模科学計算システムでの SAS の利用

大規模科学計算システムでは SAS アプリケーションは gen3 コンピューター(gen3.cc.tohoku.ac.jp)に導入されているので、そこにまずはじめにアクセスしなくてはならない。アクセスの仕方はディスプレイマネージャの利用 (X Windows環境) の方法と非対話モードでの実行がある。前者では、まず、X Window システム環境の設定をします。次に SAS を起動するには

2.1.1) gen3% sas

と入力すると図1から図4までの"LOG","OUTPUT","PROGRAM EDITOR","TOOL BOX"の4つのウィンドウが表示されます。

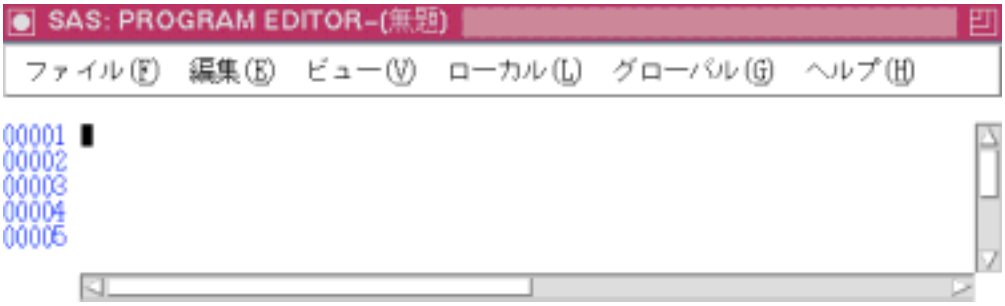

図1　プログラムエディタのウィンドウ。ここに SAS プログラムを書く

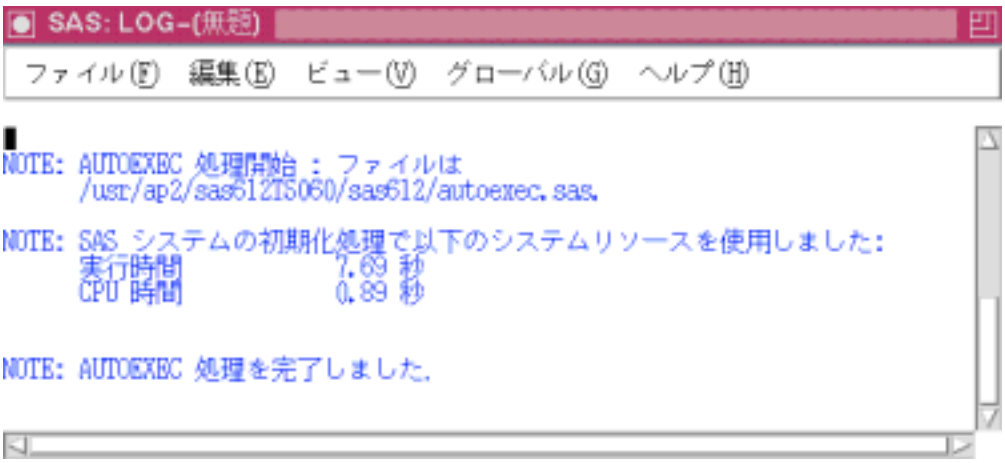

図2　. Log ウィンドウ。ここにコンピューターから実行経緯が示され、エラーやワーニングが書かれる。

図3. アウトプットウィンドウ。ここに計算結果が書かれる。(今は結果無しの状態)

図4. ツールボックス。実行やその他の命令がボタン形式でできる。

2.1.2) "PROGRAM EDITOR"に SAS プログラムを書き、実行すると"LOG"画面に実行経緯が示され、エラーやワーニングがあると、それぞれ赤色や青色で示される。
2.1.3) 計算結果にエラーがなければ、"OUTPUT"ウィンドウに計算結果は示される。
2.1.4) エラーが出た場合、プログラムエディターで編集して、正しい答えが出るまで実行を繰り返す。

後者では X Window 環境でない利用であり、Telnet で gen3 にアクセスして計算する方法である。それは以下の方法でアクセスする。

2.2.1) FTP で前もって SAS プログラム (拡張子が .sas) とデータファイル (拡張子は .txt など) を gen3 に送っておく。
2.2.2) Telnet で gen3 にアクセスする。
2.2.3) たとえば SAS プログラム名が test01.sas だとすると、
gen3% **sas test01** と入力する。
2.2.4) 実行が終わると、 test01.log と test01.lst ファイルが作られる。
2.2.5) .log には実行経緯が .lst には計算結果が書き込まれている。
2.2.6) これらのファイルを自分の PC にダウンロードして、結果を見る。

3) SAS のデータの入力

SAS のデータファイルは様々な記述形式に適応するようになっているが、ここでは一般的な方法について説明する。SAS にはそれぞれの統計分析を行うためのプロシージャ (計算手順) がすでに用意されているので、簡単な命令文の形式 (プログラムの書き方) を理解できればすぐにこれらの分析を行うことができる。SAS のプログラムでデータの読み込みは、後にどのような分析を行うにしても、どのデータファイルを用い、そのデータがどのように入力されているのかを指定しなければならない。

「eda.txt」というデータファイルという形式で入力されていたとする。

個体番号	品種	性	体重	体長
1	A	F	21.5	5.1
2	B	M	23.8	6.3
:	:	:	:	:
:	:	:	:	:

データは行列表示になっており、行は観測値（Observation）、列は変量（Variable）といっている。「データの区切り（デリミタ）は一つ以上のスペースになっていて、必ずしもカラムが揃っている必要はない。また、ここで「個体番号・・・」のような変数名の行は、読み込みの前に消去するか、書かれていても、FIRSTOBS=n のデータセットオプションを付けることで n 行目からデータを読み込むことを指定できる。このような変量は文字変数でも数値変数でも良い。これらのデータを SAS で読み込む SAS プログラムは次のようである。

注意！！　これ以降、SAS プログラムの行の説明を（数字）の後に説明する。
プログラム文は大文字でも小文字でもかまわない。
すべての行の後ろに「;」をつける。

```
DATA  dd1;                                                    (1)
INFILE 'eda.txt';                                             (2)
INPUT  NUM  BREED  $  SEX  $  WEIGHT  LENGTH;                (3)
WARIAI=WEIGHT/LENGTH;                                         (4)
OPTIONS  NOCENTER  PAGESIZE  =50  LINESIZE=100;               (5)
RUN;                                                          (6)
```

（1）DATA 文：プログラム内でのデータ名を定義する。この場合 dd1（何でも良い）。
（2）INFILE 文：「eda.txt」というデータファイルから読み込む。
（3）INPUT文：変数並びの定義。左から、個体番号、品種、性、体重、体長の順で入力されている。品種と性は数字ではなく文字変数なので、**文字変数の後ろに「＄」をつける。**
（4）新しい変数を作る。＋，－，＊，/　を用いて「WARIAI」という新しい変数を作ることができる。
（5）出力ファイルの 1 ページの大きさ指定するオプション文。50-100 程度が見やすい。
（6）ここで RUN 文
以下のプロシージャは上のデータ読み込みの「RUN;」の後に続けて書いていく。

4) 基本統計量

平均値、標準偏差、最小値、最大値を求める。

```
PROC MEANS;                    (1)
VAR WEIGHT LENGTH;             (2)
RUN;
```

(1) PROC 文(プロシジャー:プロック):どのような統計計算を行うのかを指定、この場合、体重と体長の平均値を求める。
(2) VAR 文:変数の指定。

5) 品種ごとの平均値を出したい場合

品種ごとに平均値を出したいときは、まず「SORT」というプロシージャを用いてデータを品種順に並び替え, その後に「MEANS」で品種ごとに体重と体長と割合について平均値を求める。

```
PROC SORT;
BY BREED;
PROC MEANS;
BY BREED;
VAR WEIGHT LENGTH WARIAI;
RUN;
```

出力結果例

N	Mean	Std Dev	Minimum	Maximum
25	1894.00	320.1597258	1380.00	2562.00

データ数　平均値　標準偏差　最小値　最大値

6) 分布の正規性や外れ値

平均値などの基本的な統計量のほかに、分布の正規性の検定や外れ値の判定なども行える。

```
proc univariate normal plot;
```

7) 頻度分布

```
PROC FREQ;                    (1)
TABLE BREED SEX;              (2)
RUN;
```

```
PROC FREQ;                    (1)
TABLE BREED*SEX;              (3)
RUN;
```

(1) 基本統計量、頻度分布プロシージャの実行。
(2) 品種と性別ごとの頻度。
(3) 品種と性別を組み合わせた水準の頻度。いわゆるクロス集計。

－出力結果 (例)
要因の組合せごとに下のような並びで値が示される。
性 1 性 2
品種 1 データ数
百分率(%)
行の百分率(%)
列の百分率(%) 26
27.66
56.52
53.06
品種 2 (省略) : :

8) 相関係数

二つの変数の間の親密の程度や関連性などを表す。グラフにプロットすると関連性を視覚的に把握することができる。また相関係数という数値で関連性の強さを表すことができる。

```
PROC CORR;                    (1)
VAR WEIGHT LENGTH;            (2)
RUN;
```

(1) 相関係数行列を打ち出す。相関がないことに対する有意性検定の確率が表示される。
(2) 変数を指定。三つ以上だと行列表示になる。

9) 散布図

```
PROC  PLOT;                          (1)
PLOT  WEIGHT*LENGTH;                 (2)
RUN;
```

(1) 2 つの変数の散布図を描く。
(2) 変数を指定。

```
PROC  PLOT;                          (1)
PLOT  WEIGHT*LENGTH=BREED;           (3)
RUN;
```

(3)「品種ごと」という指定をしたいとき、このように書くと、品種ごとにグラフ化される。
(2) の場合はすべての品種が同時にプロットされる。

10) 回帰

目的変数 (Y) が説明変数 (X) に依存する関係を調べ、XからYを推定することである。今、XがYに直線的、または曲線的に回帰していると想定し、Y=b0+b1*X (1 次回帰、直線) または Y=b0+b1*X+b2*X2 (2 次回帰、曲線) とおいて、b0～b2 を推定する。定数b0～b2を求めるには、最小自乗法を用いてyの実際の値と予測値との差が最も小さくなるように計算する。

たとえば、今までのデータに、AGE (日齢) という変数がさらにあったとし、日齢に対する体重の直線および曲線回帰を求める。

```
DATA;                                (1)
INFILE  'ファイル名';                (2)
INPUT     ～      AGE;               (3)
AGE2=AGE*AGE;                        (4)
PROC  REG;                           (5)
MODEL  WEIGHT  =  AGE;               (6)
MODEL  LENGTH  =  AGE2  AGE;         (7)
RUN;
```

(1) ～ (3) 今までと同じ。データの読み込み。
(4) 曲線回帰を調べるために AGE2 (二乗の変数) という値を作っておく。

(5) 回帰のプロシジャー。
(6) 直線回帰　Y=b0＋b1*AGE
(7) 曲線回帰　Y=b0＋b1* AGE＋b2*AGE2

出力結果例

Analysis of Variance (モデル式の検定)

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
	モデルの自由度	偏差平方和	平均平方	F値	P値(※1)
Model	2	39031830.15	19515915.075	208.940	0.0001
Error	91	8499784.1589	93404.221527		
C Total	93	47531614.309			

Root MSE	305.62104	R-square (寄与率)	0.8212(※2)
Dep Mean	1360.22340	Adj R-sq	0.8172
C.V.	22.46844		

※1
P 値とはこのモデルが適合していないという仮説を F 検定で検定したときの仮説の確からしさを表し、これがある基準 (0.05 または 0.01) よりも低いということは、このモデル式が適合しているということがいえる。
※2
R-squre は寄与率または決定係数と呼ばれ、目的変数の分散とモデルによって説明される分散の比率を表し、0～1 の値をとる。これが 1 に近いということは、モデル式によってより多くの分散が説明できる、つまりモデルの適合度を表してる。

Parameter Estimates (係数の予測値と検定)

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
	自由度	※3	標準誤差	※4	※5
INTERCEP	1	-608.127282	172.97827755	-3.516	0.0007
AGE2	1	-5.801246	1.57686842	-3.679	0.0004

AGE　　1　　277.585313　　35.18249136　　7.890　　0.0001

※3

この値がそれぞれの変数の回帰係数を表している。INTERCEP は指定した変数以外の誤差、つまり直線、曲線回帰の定数項である。

このモデルでは Y = - 5.80 * AGE2 + 277.58 * AGE - 608.13 という式が立てられた。

※4

それぞれの変数の回帰係数が 0 であるという仮説をt検定している。

※5

t検定での仮説の確からしさを表し、これがある基準よりも低いということは、その説明変数が目的変数へ有意な影響を与えていると言える。

11) t検定

ある 2 つの集団の平均値の差が統計的に意味のあるものなのか、偶然によって生じた程度のものなのかを判定する。検定を行うにはまず仮説を立て、その仮説がなりたつ確率を計算する。t検定では「2 つの集団の平均値には差がない」という仮説を立て、これを「H:0」(帰無仮説) で表す。

H:0 が成り立つとき「t＝2 群の平均値間差／2 群の平均値間差の標準偏差」というt値は、t 分布 (差がないという仮説のもとでのt値の理論分布) に従う。得られたt値より大きい値が得られる確率が p 値 (Prob > ｜T｜ で表す) である。p 値がある基準より小さいということは成り立たない、ということを表し、H:0 は棄却される。結論として 2 群の平均値間には有意な差があるということができる。

H:0 を棄却するかどうかのp値の基準には 0.05 や 0.01, 0.001 をよく用いる。もし「Prob>｜T｜」が 0.05 より小さければ、「2 群の平均値間は危険率 5%で有意に差がある」というういい方をする。

```
PROC TTEST;                    (1)
CLASS BREED;                   (2)
VAR WEIGHT LENGTH;             (3)
PROC TTEST;                    (1)
CLASS SEX;                     (4)
VAR WEIGHT LENGTH;             (3)
RUN;
```

これをまとめて

```
CLASS BREED SEX;
VAR WEIGHT LENGTH;
RUN;
```

とはできない。

(1) t テストを行え。
(2) 品種ごとにクラス分けして。
(3) WEIGHT と LENGTH の2変数で。
(4) 性ごとに分けて。

出力結果例

	N	Mean	Std Dev	Std Error	Variances	T	DF	Prob>|T|
	データ数	平均値	標準偏差	標準誤差	分散(※1)	t値	自由度	※2
C	48	12.25000000	3.54604965	0.51182818	Unequal	10.4223	74.3	0.0001
N	46	6.10869565	1.98021616	0.29196694	Equal	10.3054	92.0	0.0000

For H:0 Variances are equal, F' = 3.21 DF = (47,45) Prob>F' = 0.0001 (※3)

※ 1
t検定のp値 (Prob>|T|) は 2 群の分散が等しくないとき(Unequal) と等しいとき (Equal) にそれぞれ異なる方法で行われた検定の結果が表示されている。どちらを用いるかは※3 に表示されているF検定 ((8)で述べる) の結果を見て分散が等しいかどうかを確認し Unequal と Equal のどちらかのp値を採用する。F検定では「Prob>F'」がある基準よりも小さければ分散が等しくないとみなせるので Unequal を、ある基準より大きければ Equal のp値 (Prob>|T|) を用いる。この場合は 0.01 よりも小さいので分散は等しくないとみなし、Unequal を採用する。
※ 2
平均値は等しくないという仮説が起こる確率が 0.0001 とかなり低いので、両群の平均値は0.1%有意で異なるということができる。

12) F検定

ある 2 つの集団の分散が等しいと言えるか、またはその差が誤差の範囲かどうかを判定する。F検定では「H:0 分散は等しい」という仮説を立てる。

いまAとBという二つの集団があるとき、「F=Aの分散／Bの分散」というF値を計算すると、H:0 が成り立つときこの値はF分布 (H:0 という仮定の下でのF値の理論分布) に従う。

得られたF値より大きい値が得られる確率がp値 (Prob＜F で表す) であり、p値がある基準よりも小さいということは H:0 は成り立たないということを表し、H:0 は棄却される。結論として

分散は等しくない、ということができる。

13) GLM（一般線形モデル）による多変量分散分析

分類変数（例：品種、性）を説明変数とした線形モデルを作り、分散を分解することでそれぞれの説明変数が目的変数（例：体重）に及ぼす効果を調べる。効果の有無の判定はF検定を用いる。

今、ある集団の体重は品種と性の影響を受けていると仮定して

体重＝全平均＋品種の影響＋性の影響 ＋エラー

$Y_{ijk}=\mu+B_i+S_j+\varepsilon_{ijk}$

ここで、μ：全平均

B_i：品種の効果

S_j：性の効果

ε_{ijk}：誤差（未知の効果）

という線形モデルを設定する時、全分散も次のようにそれぞれの効果に分解することができる。

全分散＝品種による分散＋性による分散＋誤差の分散

次に「H:0 品種による分散と誤差の分散が等しい」という仮説をたて、F検定を行う。

検定の結果、「Prob>F」が基準より小さいということは H:0 は成り立たない、つまり品種と誤差の分散は等しくないので、品種の影響の大きさは誤差の影響の大きさと有意に異なる。よって品種は体重に有意な影響を与えているということができる。

```
PROC GLM;                                          (1)
CLASS YEAR PLACE;                                  (2)
MODEL WEIGHT LENGTH = YEAR PLACE SIRE;             (3)
RUN;
```

(1) GLM 分析を行え。

(2) 説明変数は YEAR と PLACE である。

(3) 一般線形モデルは従属変数が WEIGHT,LENGTH で説明変数は YEAR,PLASE,SIRE である。目的変数が「＝」の左、説明変数を右に書く。

この時、SIRE は CLASS に現れていないので変量効果とみなされる。また、MODEL の説明変数の欄に「YEAR*PLACE」を加えれば YEAR と PLACE の交互作用、「YEAR PLACE(YEAR)」とすると、PLACE は YEAR の枝分かれ構造であることを表す。

出力結果例

Dependent Variable: EDA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	40938037.2	10234509.3	138.15	0.0001
Error	89	6593577.0	74085.1		
Corrected Total	93	47531614.3			

R-Square	C.V.	Root MSE	EDA Mean
0.861280	20.01038	272.18584641	1360.22340426

回帰分析を同じように、まずそのモデルのF検定の結果が出力される。「Pr > F」が基準以下ならばそのモデルは適合しているといえ、その適合の度合いは決定係数(R-Square)で示されている。この値が低い時は、CLASS で取り上げた要因以外に目的変数に影響を与えている要因が存在することを意味する。

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BREED	1	38571786.7	38571786.7	520.64	0.0001
SEX	1	660618.9	660618.9	8.92	0.0036
PRO	1	188.3	188.3	0.00	0.9599
ENERGY	1	1705443.1	1705443.1	23.02	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BREED	1	39466225.8	39466225.8	532.71	0.0001
SEX	1	810531.2	810531.2	10.94	0.0014
PRO	1	6142.3	6142.3	0.08	0.7741
ENERGY	1	1705443.1	1705443.1	23.02	0.0001

それぞれの説明変数の目的変数に対する効果の有意性は「Type I SS」と「Type III SS」の2 種類出力される。Type I は各説明変数ごとのデータ数がすべて同じときに用いることができ、データ数が異なるときは Type III を用いる。

それぞれの説明変数のF値(Pr > F)が基準以下なら、その説明変数は目的変数に有意な影響を与えていると言える。この例では「品種とエネルギーの効果は危険率0.1%、性は危険率 1%でそれぞれ有意な影響を与えており、たんぱく質は有意な影響を与えていなかった」ということができる。

文献

1) 小野敏 「ライブラリー・アプリケーションソフトウエアーの紹介」 SENAC Vol.35,No.1,P42-44(2002,4)

2) SAS 出版局 「Base SAS ソフトウエアー使用法ガイド」

3) SAS 出版局 「SAS /STAT ソフトウエア ユーザーズガイド」

4) 竹内啓ら 「SAS によるデータ解析入門」 東京大学出版会