SAS
ユーザー総会
2023

SASユーザー総会2023論文集

会場：東京大学大学院 情報学環・福武ホール

会期：2023年9月13日(水) ～14日(木)

主催：SASユーザー会 世話人会

目次

論文部門

抄録----プレゼンテーション部門

セルオートマトン・シミュレーションのSAS実装

森岡　裕

（イーピーエス株式会社）

Cellular automata in SAS

Yutaka Morioka

(EPS Corporation)

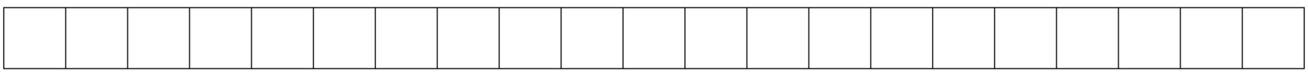
要旨

セルオートマトンは，離散的な時間区切りを設定し，ある状態をもつ，ある時点の区画(セル)の次時点の状態が，セル同士の相互の関係によって決定される集合を用いたシミュレーションの一種である．一般的に解析プログラム言語を習得する上で必要となる要素，データハンドリング能力・ビジュアル化のスキル・アルゴリズムの理解などの成長を総合的に促すため，教育上有用なテーマと考える．SAS コードでの実装例をまじえ，1 次元の単純な定義から順番に紹介していきたい．あくまで，セルオートマトンそのもの理論的説明が主題ではなく，統計解析プログラムや SAS に興味を持つためのツールとして，セルオートマトンを解説したい．

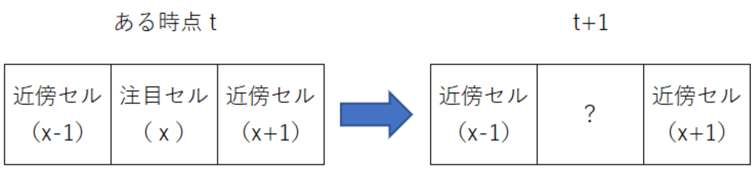
キーワード：セルオートマトン

1 次元セルオートマトン

例えば，四角形で区切られたセルというものが延々に続いているイメージを図で描いてみる．



そこから任意のセルを対象として取り出し（まとまりを近傍とも表現する），「注目セル」x を定義する．
隣り合ったセルを「近傍セル」と定義し，1 つ左隣りのものを x-1，右隣のものを x+1 と表現する．



ある時点を t とした時，t における注目セル x の次の時点での状態，つまり t+1 時点での状態が，t 時点の x の値および x-1，x+1 の値によって決定されると設定する．

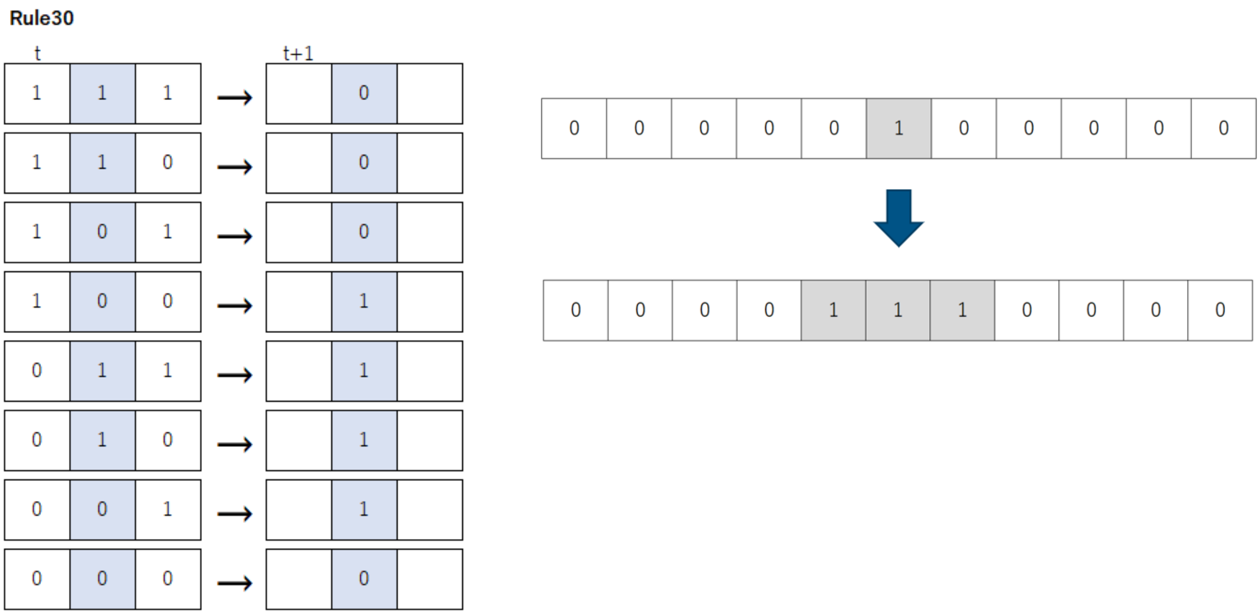
近傍を 3，セルが取りうる値を 0 か 1 とすると，パターンとしては(0，0，0)から(1，1，1)までとなり $2^3=8$ 通りとなり，決定するためのルールはそれらパターンについて，そのセルが次世代に 1 と 0 のどちら状態とな

るかを決めるため，2^8=256 通り存在することになる．

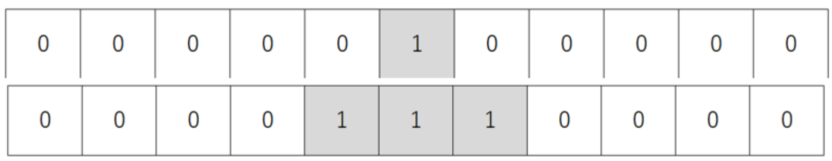
それら 3 近傍セット 2 値における 256 のルールパターンはすでに，挙動が研究されている．その中から代表的なものを紹介する．

Rule 30

3 近傍セットで状態 2 値の一次元セルオートマトンのうち，有名なものをとりあげる．Rule30 と呼ばれるもので，左下図の左側が，時点 t の状態のパターンと注目セルの値であり，右側が t+1 の際の注目セルの値になる．t+1 の注目セルを上から縦に 00011110 と読むとこれを 2 進数と考えて，10 進数に変換すると 30 になるのが Rule 30 と言われる所以である [1]．左下図のルールを見た上で，右側の状態遷移の例をみると，なぜ値 1 のセルが 1 つから 3 つに拡大する挙動が理解できるはずである．



さて，この Rule30 の遷移結果を，t から t+n まで繰り返し，結果を積み木のように重ねていくとどうなるであろうか.




```
data wk1;
array col{200}  (200*0);
array ncol{200}  (200*0);
col{100}=1;
do obs = 1 to 100;
    do i = 1 to 200;
    if obs = 1 then do;
        ncol{i}=col{i} ;
    end;
    else do;
      if i=1 then pattern=cats(0,  col{i}, col{i+1});
```

```
12          if (1+1)<=i<=(200-1) then    pattern=cats(col{i-1}，col{i}，col{i+1});
13          if i=200 then pattern=cats(col{i-1}，   col{i}，0);
14               /*rule*/
15                 select(pattern);
16                   when("111") ncol{i} =0;
17                   when("110") ncol{i} =0;
18                   when("101") ncol{i} =0;
19                   when("100") ncol{i} =1;
20                   when("011") ncol{i} =1;
21                   when("010") ncol{i} =1;
22                   when("001") ncol{i} =1;
23                   when("000") ncol{i} =0;
24               end;
25          end;
26        end;
27      output;
28      do i = 1 to 200;
29        col{i}=ncol{i} ;
30      end;
31  end;
32  keep ncol:;
33  run;
```

LINE 02 で任意の要素数の 1 次元配列を作り，同様素数で t+1 の値を格納する 1 次元配列を作っている

注目セルをループでずらしていきながら，{i-1}，{i}，{i+1}の関係性をとらえ，LINE 16-24 でそのまま Rule を select ステートメントで表現している．1 次元配列の基本的な使い方と，select による条件分岐のみで表現されており，初学者でも内容をつかみやすい．

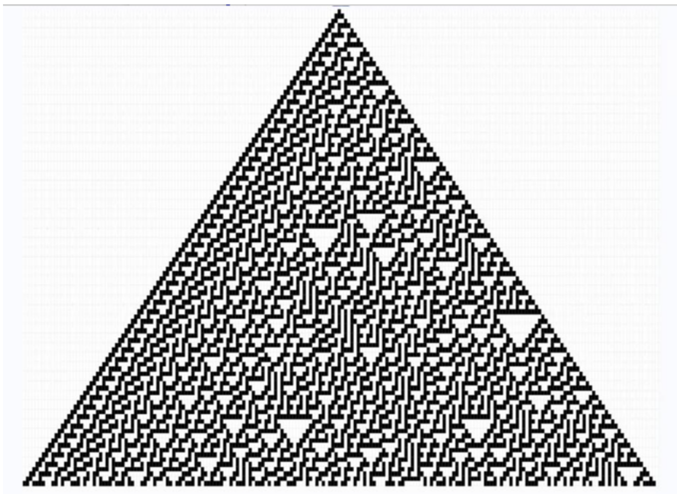
可視化の部分は SAS RWI(Report Writing Interface)を活用する

```
01  data _null_;
02  set wk1 end=eof;
03  array AR ncol:;
04  if _N_=1 then do;
05        declare odsout ob();
06        ob.table_start();
07  end;
08  ob.row_start();
09  do over AR;
10        if AR=1 then do; background="black";color="black";end;
11        else if    AR=0 then do; background="white";color="white";end;
12        text=catx(" ", " color=", color, " height=0.01 width=0.01    vjust=center background=", background);
13        ob.format_cell(data:AR, style_attr:text);
14        call missing(of background color);
15  end;
16  ob.row_end();
17  if eof then do;
18        ob.table_end();
19  end;
20  run;
```

SAS RWI は，データステップでビジュアルを作成することができる非常に強力な機能である．

LINE02 で declare を使って宣言しているところからもわかるが，RWI は SAS ハッシュオブジェクトと同じ，オブジェクトとメソッドで表現するタイプの文法をもっている．そのため，RWI から，レポート出力・ビジュアル化・ハッシュオブジェクトの操作感覚をつかむことができ，その点からも教育効果が極めて高い．

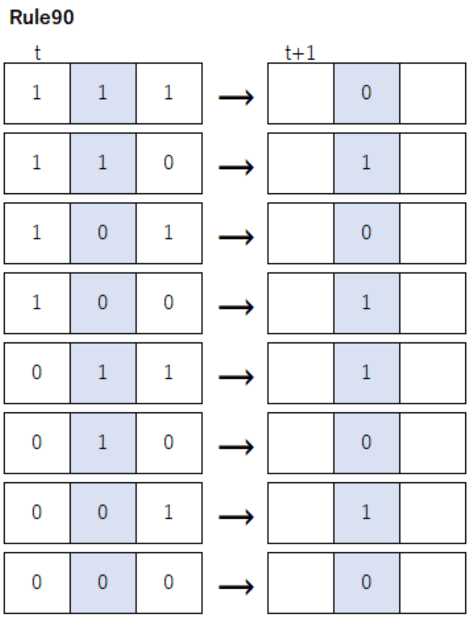
さて，では 100 時点を経過して，積み重なったセルオートマトンがどうなっているかをみてみる．



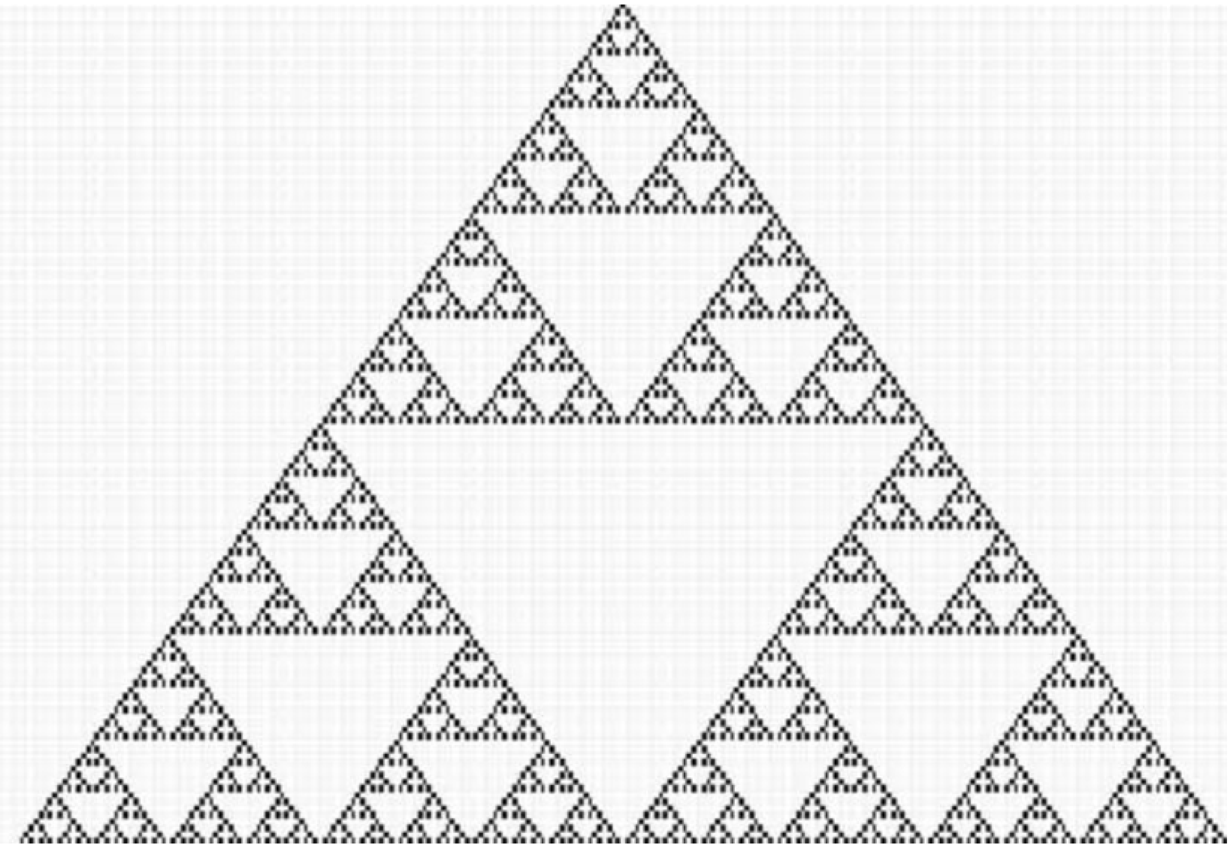
左上図が SAS の RWI の出力である．そして右上画像はイモガイの一種タガヤサンミナシの貝殻の紋様になる[2]．Rule30 が紋様に類似することは既に研究されており，生物における「各色素細胞は活性化されると色素を分泌し，同時に近傍の色素細胞の活性化に影響を及ぼす」という特性が，セルオートマトンのアルゴリズムと部分的に一致しているからだとみられている．
プログラムによるアルゴリズム実装とビジュアル化が，自然界の現象とリンクする不思議である．この，プログラムが動き，画像が生成されて，開発者が「不思議」を感じるという一連の流れが，プログラマの好奇心と学習意欲をこれ以上なく刺激するスパイスになると主張したい．

Rule 90

フラクタル(fractal)とは,「自己相似性」ともいわれる幾何学的性質で，部分のパターンと全体のパターンが相互に再現されているような状態である．画家ゴッホの作品の中にも，フラクタルを巧みに取り入れたものが存在することが有名である．
Rule 90 はフラクタル図形として有名なシェルピンスキーのギャスケットを描くことができる．先のコードの select ステートメントを貼り換えるだけで完成となる．




```
select(pattern);
  when("111") ncol{i} =0;
  when("110") ncol{i} =1;
  when("101") ncol{i} =0;
  when("100") ncol{i} =1;
  when("011") ncol{i} =1;
  when("010") ncol{i} =0;
  when("001") ncol{i} =1;
  when("000") ncol{i} =0;
end;
```



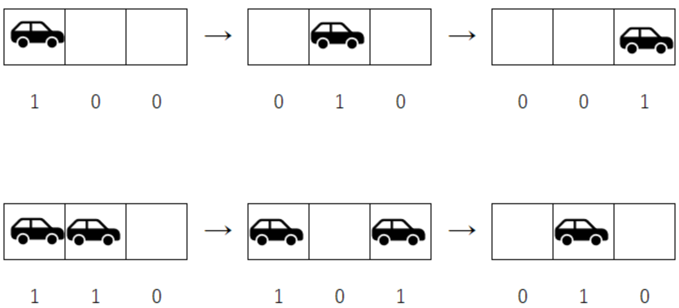
8 通りの単純なルールパターンでこのような美しい図形を描けることには驚きを生じさせる

Traffic Rule (Rule 184)

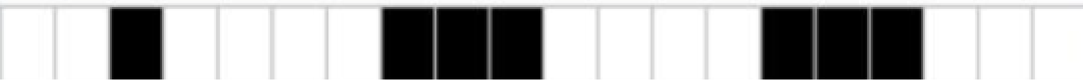
渋滞学という学問分野において，最も基礎モデルとして紹介されるものに Rule 184 があり，Traffic Rule とも呼ばれる．[3]

Rule184

t				t+1		
1	1	1	→		1	
1	1	0	→		0	
1	0	1	→		1	
1	0	0	→		1	
0	1	1	→		1	
0	1	0	→		0	
0	0	1	→		0	
0	0	0	→		0	



ルールとして，上右図のように，目の前のセルが 0 なら 1 が進み，目の前のセルが 1 なら待機とみなせるような挙動をとる．これは俯瞰的にみれば，車間距離が詰まれば停滞し，空けば進むという車両交通の挙動を示すことになる．
以下は部分的な小渋滞が解消していく様子を示している．



↓

↓

論文上で示すことが難しいが，この Traffic Rule は画像をコマ送りにする GIF アニメーションなどで表現すると非常にわかりやすい．
あらかじめ，Traffic Rule で，累積のデータセットを作っておき，

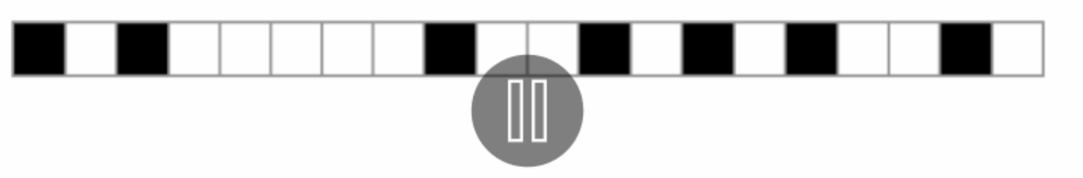
```
data wk1;
array col{20}    (20*0);
array ncol{20}    (20*0);

col{3}=1;
col{8}=1;
col{9}=1;
col{10}=1;
col{15}=1;
col{16}=1;
col{17}=1;
do obs = 1 to 50;
        do i = 1 to 20;
        if obs = 1 then do;
                        ncol{i}=col{i} ;
        end;
        else do;
          if i=1 then pattern=cats(col{20},    col{i}, col{i+1});
          if (1+1)<=i<=(20-1) then    pattern=cats(col{i-1}, col{i}, col{i+1});
          if i=20 then pattern=cats(col{i-1},    col{i}, col{1});
                        /*rule*/
                          select(pattern);
                            when("111") ncol{i} =1;
                            when("110") ncol{i} =0;
                            when("101") ncol{i} =1;
                            when("100") ncol{i} =1;
                            when("011") ncol{i} =1;
                            when("010") ncol{i} =0;
                            when("001") ncol{i} =0;
                            when("000") ncol{i} =0;
                        end;
          end;
        end;
    output;
    do i = 1 to 20;
        col{i}=ncol{i} ;
    end;
```

```
end;
keep ncol:;
run;
```

それを loop で一つずつ抽出し，RWI で描画する処理について，opitoons animation=start でアニメーション化の開始点設定，animduration=xx でコマ送りの間隔を設定し，printerpath=GIF を指定（下図コード LINE05）する．

```
01 ods _all_ close;
02 ods listing;
03 /*GIF アニメの開始設定*/
04 ods graphics   / width=220mm height=20mm ;
05 options   nodate animation=start   animduration=0.5     printerpath=GIF papersize = ("350mm"， "50mm") ;
06 ods printer file="XXXX¥trafic.gif";
07 %macro loop;
08 %do i = 1 %to 20;
09 data _null_;
10 set wk1 ;
11 where monotonic()=&i;
12 array AR ncol:;
13 dcl odsout ob();
14 ob.table_start();
15 ob.row_start();
16 do over AR;
17         if AR=1 then do; background="black";color="black";end;
18         else if   AR=0 then do; background="white";color="white";end;
19         text=catx(" "， " color="， color， " height=10mm width=10mm   vjust=center background="， background);
20         ob.format_cell(data:AR， style_attr:text);
21         call missing(of background color);
22 end;
23 ob.row_end();
24 ob.table_end();
25 run;
26 %end;
27 %mend;
28 %loop;
29 /*アニメ終了*/
30 options   ANIMATION=STOP ;
31 ods printer close ;
```

LINE31 で ods printer close するまでの出力が含まれる形で GIF ファイルの作成が完成する．



Traffic ルールは，自由にルールを改造することができ，車間距離の検知を 2 セル先までに変更したり，車線を追加して 2 次元にし，前が詰まっていて，隣の車線があいていればそちらに車線変更といったアルゴリズムも実装可能である．渋滞学の基礎モデルと言われる所以はそういった性質からである．

2 次元セルオートマトン

2 次元セルオートマトンは，セルが方眼紙上の空間に敷き詰められているイメージとなる．近傍セルについては典型的なものが 2 種類定義されている．

次頁の左図，注目セルの縦横の隣接を近傍セルとするのがフォン・ノイマン近傍であり，右図，注目セルの縦横斜めの 8 つを近接セルとするのがムーア近傍である．

[フォン・ノイマン近傍]

	近傍セル (x,y-1)	
近傍セル (x-1,y)	注目セル (x,y)	近傍セル (x+1,y)
	近傍セル (x,y+1)	

[ムーア近傍]

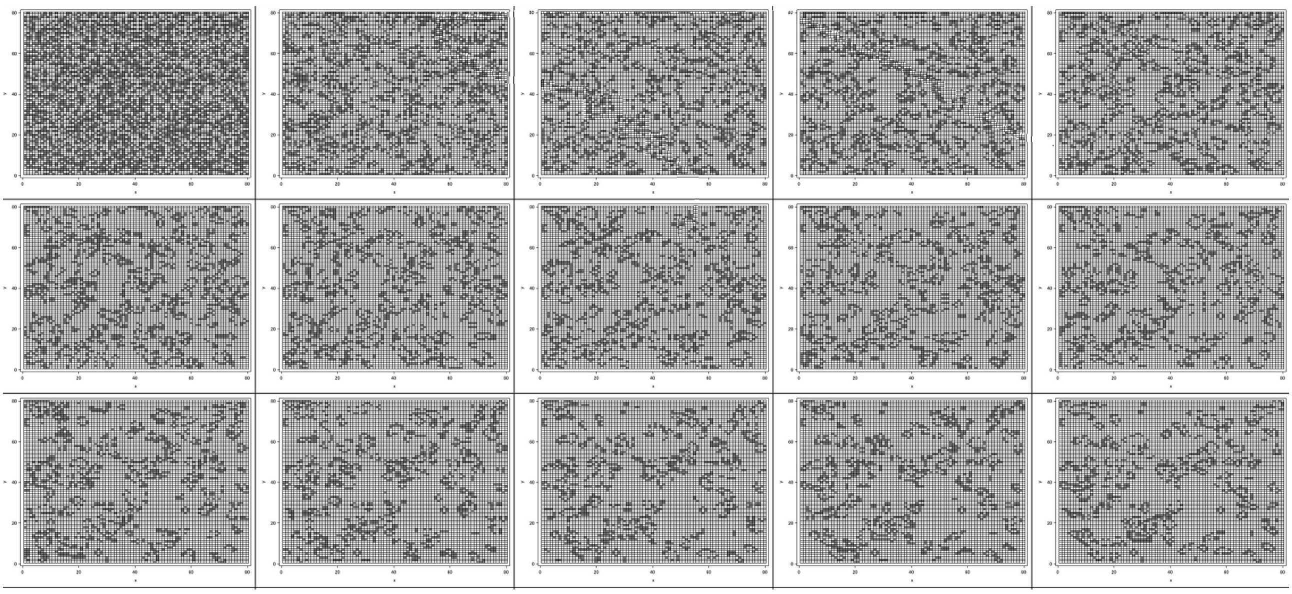
近傍セル (x-1,y-1)	近傍セル (x,y-1)	近傍セル (x+1,y-1)
近傍セル (x-1,y)	注目セル (x,y)	近傍セル (x+1,y)
近傍セル (x-1,y+1)	近傍セル (x,y+1)	近傍セル (x+1,y+1)

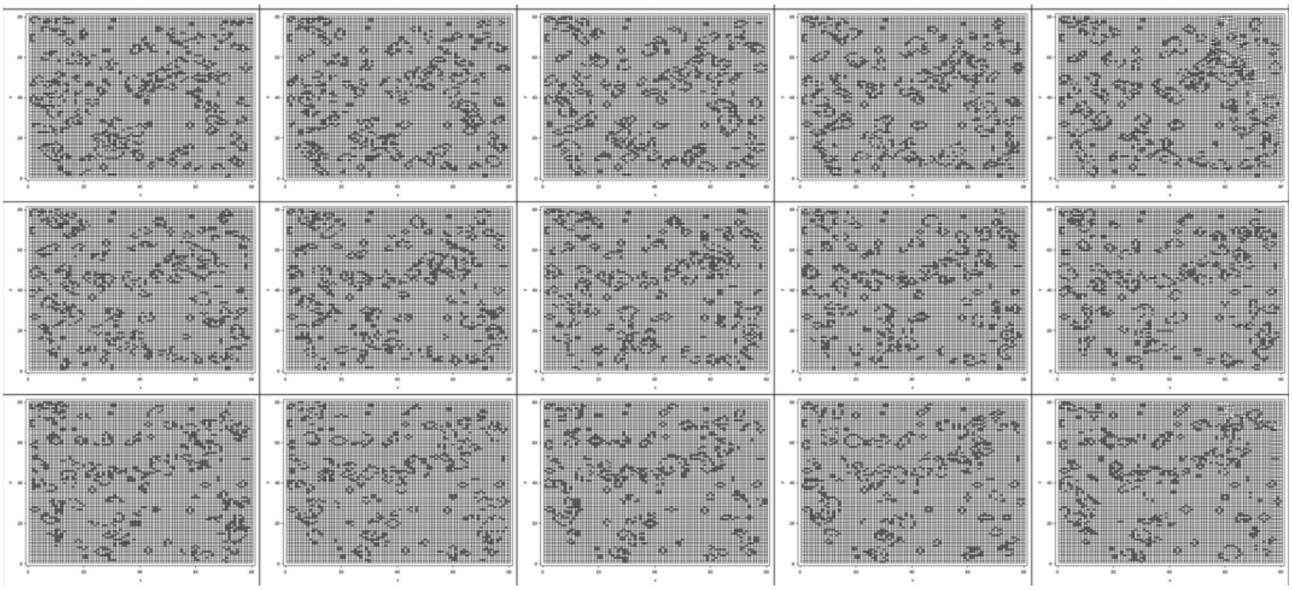
ライフゲーム

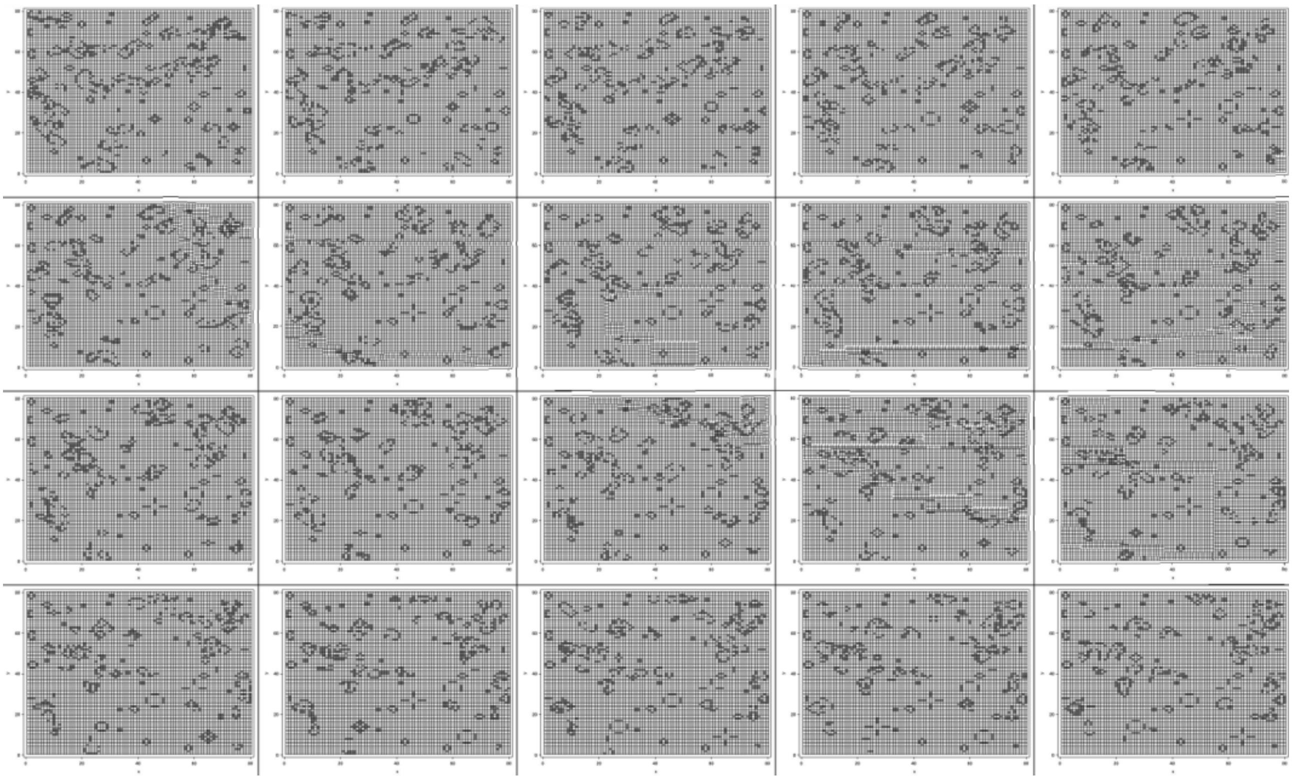
ムーア近傍で状態 2 値の 2 次元セルオートマトンとして，ジョン・ホートン・コンウェイの発明したライフゲームが有名である[4]．注目セルと近傍セルの関係性と，次時点の状態決定を生物集団の生存・死滅になぞらえて．以下のようなルールを定義したものである．

誕生：死滅セルに生存セルが 3 つ近接していれば，次の世代が誕生とみなし，注目セルは生存に転じる．
生存：生存セルに生存セルが 2 つか 3 つ近接していれば，次の世代でも生存するとして生存が維持される
死滅(過疎)：生存セルに近接する生存セルが 1 つ以下ならば，過疎により生存セルは死滅セルに転じる。
死滅(過密)：生存セルに隣接する生きたセルが 4 つ以上ならば，過密により死滅する

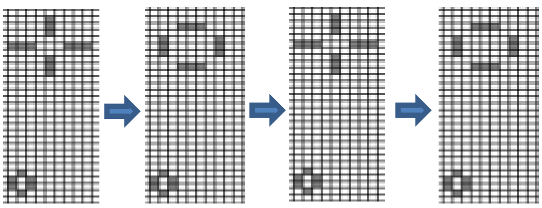
初期状態を，乱数で適当に設定し，かなり生存数が過密な状態から t=50 までとしてゲームをスタートさせてみる．死滅(過密)ルールが強く作用し，かなり疎な小コロニーのパターンにいたるまで，生存数が急減する，



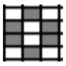
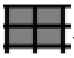


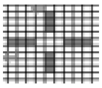
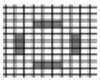
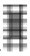


その後，局面における生存数は安定状態になり，そこから特有の周期パターンが観測できるようになる．例えば，以下は， 37 – 40 試行局面のある部分を切り取ったものだが



左下の部分は，t が進んでも形状が固定されていることがわかる．こういったパターンをライフゲームでは「固定物体」と呼ぶ．

 は蜂の巣(beehive)と呼ばれる固定物体で，生成確率は高く，ゲーム風に表現するなら出現レア度は低く，もっともよくでる最低レア度のブロック(block)の次にありふれたパターンである.

とが繰り返しているように見える部分だが，これは　と　の繰り返しが 4 つ集合してる状態である. 周期性をもって，t+n 時点で t の状態に戻るものをライフゲームでは「振動子」と呼んでおり，3 セルの｜と━の交互パターンは，ブリンカー(blinker)と呼ばれ，振動子の中ではもっともレア度の低い．こちらも，ありふれたパターンである.

実はライフゲームでは，初期配置によっては，途方もない周期性かつ巨大な振動子をつくることが可能であり，単純なルールで生成されているとは信じられないようなパターンが数多く発見されている.

これも，論文では表現ができないが，アニメーションにして考察するべきセルオートマトンである.

SAS での実装は SAS IML 製品ライセンスをつかわないと，一見難易度が高いかのように思えるが，実質 100 行もいかない程度の平易なコードで実装可能である．ポイントとなるのは SAS ハッシュオブジェクトによる自己参照で，set するのと同じデータセットをハッシュオブジェクトに格納して x 座標と y 座標を key にして，値を data にしてしまえば，後はループしながら find メソッドで(y-1，x-1)〜(y+1，x+1)までのムーア近傍にあたる値をとって，近傍の生存数をカウントすればいいだけなので，初歩的な問題に落とし込むことができる.

```
data wk1;
call streaminit(777);
do x=1 to 80;
  do y=1 to 80;
    if rand("uniform")<0.5 then v=1;
    else v=0;
    output;
  end;
end;
run;

%macro calc;
data wk2;
  set wk2;
    if _N_=1 then do;
      call missing(of　cx cy cv);
      declare hash h1(dataset:"wk2(rename=(x=cx y=cy v=cv))");
      h1.definekey("cx"，"cy");
      h1.definedata("cv"，"cx"，"cy");
      h1.definedone();
    end;
    if h1.find() ne 0　then call missing(of cx cy cv);

    cx=x;
    cy=y;


    cx=x-1;
    cy=y-1;
    if h1.find() ne 0　then xm1ym1=0;
    else xm1ym1=cv;

```

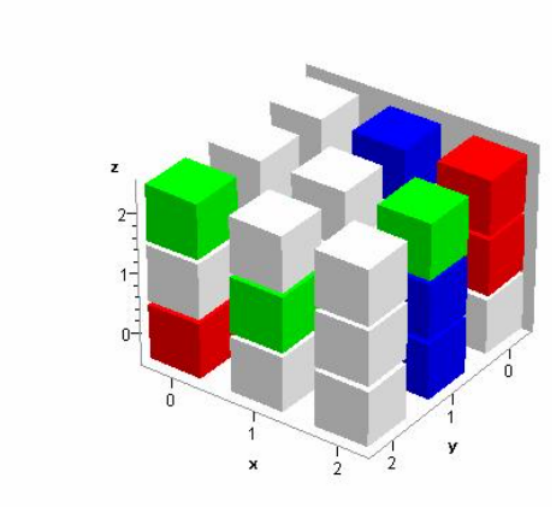
```
    cx=x;
    cy=y-1;
    if h1.find() ne 0    then x0ym1=0;
    else x0ym1=cv;

    cx=x+1;
    cy=y-1;
    if h1.find() ne 0    then xp1ym1=0;
    else xp1ym1=cv;

    cx=x-1;
    cy=y;
    if h1.find() ne 0    then xm1y0=0;
    else xm1y0=cv;

    cx=x+1;
    cy=y;
    if h1.find() ne 0    then xp1y0=0;
    else xp1y0=cv;

    cx=x-1;
    cy=y+1;
    if h1.find() ne 0    then xm1yp1=0;
    else xm1yp1=cv;

    cx=x;
    cy=y+1;
    if h1.find() ne 0    then x0yp1=0;
    else x0yp1=cv;

    cx=x+1;
    cy=y+1;
    if h1.find() ne 0    then xp1yp1=0;
    else xp1yp1=cv;

    count=sum(of xm1ym1 x0ym1 xp1ym1 xm1y0 xp1y0 xm1yp1 x0yp1 xp1yp1 );

    if v = 1 then do;
        if count in (2:3) then v = 1;
        else v = 0;
    end;
      if v = 0 then do;
        if count =3 then v = 1;
        else v = 0;
    end;

    drop cx cy    cv xm1ym1 x0ym1 xp1ym1 xm1y0 xp1y0 xm1yp1 x0yp1 xp1yp1 count;
  run;

proc sgplot data=wk2;
styleattrs datacolors=(white gray);
     heatmapparm x=x y=y colorgroup=v / outline    ;
run;
%mend;
%macro loop(times=5);
  proc sgplot data=wk1;
  styleattrs datacolors=(white gray);
      heatmapparm x=x y=y colorgroup=v / outline ;
      gradlegend;
run;
data wk2;
        set wk1;
run;
```

```
%do i=1 %to &times;
  %calc;
%end;
%mend;
ods _all_ close;
ods listing;
options   nodate animation=start   animduration=0.5    printerpath=gif   ;
ods printer file="¥test¥lifegame.gif";
%loop(times=50)
options   animation=stop ;
ods printer close ;
```

3 次元拡張

1 次元から 2 次元のセルオートマトンの実装で，データステップにおける array 配列処理とハッシュオブジェクトをマスターすることができれば，n 次元への拡張は容易である．

3 次元セルオートマトンまでで，あればビジュアル化することも平易であるが，SAS の場合，3 次元プロットの機能は貧弱といってしまってもよい状態である．SAS GRAPH の製品ライセンスは必要であるが，ほぼ裏技/チートに近い形で，3 次元セル構造をプロットする方法を発見したので，その方法論を共有したい．セルオートマトンのアルゴリズム部分は，本論文を読んでいただいた方が，好きに実装していただければと思う．




```
goptions device=javaimg;
axis1 order = (-2 to 6 by 1);
axis2 order = (-2 to 6 by 1);

%macro IMGSIZE(w=500,   h=500,   dpi=180,   rows=50,   cols=50);

    %if &dpi<=0 %then
        %put DPI must be greater than zero.;
    %else %do;
        goptions hsize=%sysevalf(&w/&dpi)in vsize=%sysevalf(&h/&dpi)in
                    hpos=&cols                    vpos=&rows;
    %end;

%mend IMGSIZE;
%IMGSIZE(w=1000,   h=1000);

```

```
17 %macro loop;
18 data wk1;
19 length color $10.;
20 do x=0 to 4;
21   do y=0 to 4;
22     do z= 0 to 4;
23         color="white";
24          if rand("uniform")<0.2 then color="blue";
25          if rand("uniform")<0.2 then color="red";
26          if rand("uniform")<0.2 then color="green";
27           x=x+rand("uniform")*10**-100;
28          output;
29     end;
30 end;
31 end;
32 run;
33 proc g3d data=wk1;
34   scatter x*y=z /shape="CUBE" size=4 noneedle color=color rotate=(10 to 350 by 34) tilt=(10 to 80 by 7)    zmin=-2 zmax=6
   xaxis=axis1 yaxis=axis2 ;
35 run;
36 %mend;
37
38
39 ods    noresults;
40 options    animation=start    animduration=0.5      printerpath=gif nodate;
41 ods printer file="XXXX ¥test.gif" ;
42   %loop
43 ods printer close ;
44 options    animation=stop ;
```

テクニックとして主張したいのが，G3D プロシジャには x と y 座標の組み合わせに対して z 座標を一意にしないと描画不能という縛りがあり，それゆえに立体セル集合構造を描くことができないのだが，LINE27 のように x に極小の誤差を与えることで，制約をすり抜けて事実上，描画ができるということである．あとは陰影の付き具合や立体表現がうまくでる device 探しになるが，device=javaimg が，実験した範囲では一番まともに立体セル構造を表現できていたので，これを採用した．

まとめ

1 次元･2 次元セルオートマトンを紹介し，最後に 3 次元セル構造プロットの独自方法論のみ記載した．セルオートマトンは，研究分野として非常に興味深いが，本稿では既に明らかになっている有名な例を紹介したのみであり，理論を進めたものではない．配列・key-data 構造の操作によって値を取得し，アルゴリズムを通して，次の値を決定，さらにそれを様々な方法でビジュアル化すること，それをプログラムで実装することによって，プログラマの練度があがるだけなく，プログラミングの楽しみを見出してもらえるのではないかというのが本稿の主張点である．SAS プログラマがより日々のコーディングを楽しんで行えるように，今後も話題を提供していきたい．

参考文献

[1] Wolfram,　Stephen (July 1983). “Statistical Mechanics of Cellular Automata”. Reviews of Modern Physics 55 (3): 601–644.

[2] Coombs,　Stephen (February 15,　2009),　The Geometry and Pigmentation of Seashells,
https://www.maths.nottingham.ac.uk/plp/pmzsc/pdfs/Seashells09.pdf　(Accessed Aug 14,　2023)

[3] 北 栄輔,　脇田 佑希子(Dec 2011 オーム社) Excel で学ぶセルオートマトン

[4] 記事「ライフゲーム」フリー百科事典　ウィキペディア日本語版
https://ja.wikipedia.org/wiki/%E3%83%A9%E3%82%A4%E3%83%95%E3%82%B2%E3%83%BC%E3%83%A0　(Accessed Aug 14,　2023)

MIプロシジャの使用方法

○小林　邦世

（イーピーエス株式会社）

How to use missing data and proc MI ?

Kuniyo Kobayashi

EPS Corporation

要旨

MAR が仮定される欠測データについては多くの手法が紹介されているが，欠測メカニズムに Missing not at Random(MNAR)が仮定される欠測データの取り扱いについて，SAS での実装法についてまとめられている資料は必ずしも多いとはいえない．そこで本発表では MNAR が仮定される欠測データの基礎的な取り扱いと MI プロシジャを利用した多重補完法の実装例を紹介したい．

キーワード：欠測データ，MNAR, MI プロシジャ

1，緒言

昨今の臨床試験において欠測データの取り扱いについては多くの手法が提案され、数多の議論が交わされている．本論文では改めて欠測メカニズムと欠測パターンについてまとめた上で，欠測メカニズムにMNARが仮定される経時的測定データの多重補完法について，MIプロシジャを用いた基礎的な方法と実装例を紹介したい．

2，欠測について

被験者 1 人あたり，複数時点の応答変数（経時的測定データ）を持つ臨床試験において，試験薬の効果が小さい被験者の方が，より多く試験を途中で中止してしまい，計画された最終時点での欠測率が高くなった状況を考える．このとき，最終時点での値が観測された被験者のみで，応答変数の算術平均による推定を行うと，試験薬の効果が大きくなる方向にバイアスが生じる可能性がある．このように，「どのようなデータが欠測しやすいか」，すなわち欠測のメカニズムとパターンの情報が，欠測のあるデータの解析の妥当性を評価する際に重要になり得る．

本論文では，欠測について，そのメカニズムとパターンを説明したのち，Missing not at Random（データが欠損データに依存して欠損する）の場合における「多重補完法（多重代入法：Multiple Imputation）」と呼ばれる

統計手法を簡単に紹介していきたい.

3, 欠測メカニズムについて

欠測メカニズムは大きく分けて，3種類存在する[1].

Missing Completely at Random (MCAR)は，データ内のある変数が欠損する確率と他の変数との間に関係性がないタイプの欠測メカニズムのことである．臨床試験においては転居など，主要評価項目・有害事象の発現等とは全く無関係な治験中止による欠測が該当する．データが MCAR であると仮定した場合，欠損していないデータだけを使用し，分析することでバイアスの少ない結果を得ることが可能である.

Missing at Random (MAR)とは，生じている欠損のパターンを他の変数の情報から規則的に予測できる，観測されている変数で説明することができるタイプの欠損メカニズムのことである．臨床試験においては原疾患の悪化，有害事象の発現等と関係しうる理由による欠測の場合などが考えられ，特に，症状が悪化している症例が十分にある場合，もしくは，検査結果等が開示された場合に発生する欠測が該当する.

Missing not at Random (MNAR) は，データが未知の原因により欠損しているか，あるいは未知の規則に基づいて欠損するメカニズムである．基本的に他の情報に基づいて欠測を予測することは困難である．臨床試験においては，主要評価項目・有害事象の発現等と関係し得る理由による欠測であり，来院間隔が長い場合などに，欠測の原因となったデータが得られていない場合に該当する．MNAR は複雑で，取り扱いが難しいため，データ収集段階から発生を可能な限り防ぐ，あるいは、欠損が生じた原因の情報を併せて収集することが望ましいとされる[2].

臨床試験において，欠測のメカニズムが MCAR，MAR，MNAR のどれに該当するかは対象の疾患の特性・プロトコルの状況・組み入れられた被験者の状態等で異なり，試験ごとに個別の検討が必要である．欠測メカニズムによって，得られた推定量の妥当性が変わることがあり得るため，中止理由の情報を集めておくことは，当該試験の解析のみならず，類似の薬剤・疾患の治験を計画する際にも大変重要な情報となる．また，症例毎の欠測メカニズムと試験全体の欠測メカニズムを考えると，原理的には,

・ 全中止例のうち，全例 MCAR ならば，試験全体で MCAR.

・ 全中止例のうち，1 例以上 MAR が存在し，残りの全症例が MCAR ならば，試験全体では MAR.

・ 全中止例のうち，1 例以上 MNAR が存在すれば試験全体では MNAR.

とみなされることが多い[2].

4，欠測のパターンについて

欠測のパターンの重要な区分として，次の2つがある．臨床試験における被験者の脱落のように，一度データが欠測した場合，その後のデータが全て欠測となるような場合の欠測を「単調な欠測(monotone missing)」と呼ぶ．一方，ある測定時点のみ来院できなかった，ある時点のみ服薬できなかったという場合のように，一旦欠測した場合でもその後，データが 1 時点でも観測された場合の欠測を「非単調な欠測(non-monotone missing)」と呼ぶ[2](図1参照)．

被験者番号	群	Week 0	Week 2	Week 3	Week 4
001	P	X	X	X	X
002	A	X	X		X
003	P	X		X	X
004	A	X			
005	P	X	X		
006	A	X	X	X	X
……					

非単調な欠測
(Non-monotone missing)

単調な欠測
(monotone missing)

図 1　欠測を含む連続量経時測定データの非単調/単調な欠測の図

5，欠測値の取り扱い方法

臨床試験においては,試験の立案段階・プロトコル作成段階において，
なるべく欠測が発生しないような試験計画を作成することが好ましいとされている．しかしながら試験実施に際してデータに欠測が発生するのはある程度やむを得ない．そこで，欠測が発生した場合に以下のような解析手法が提案されている[3]．

・欠損値を持った症例を除いた解析方法

・欠損値のみを除いた解析方法

・欠損値を補完する解析方法

このうち，欠測値を補完する解析方法は，さらに単一補完法と多重補完法に分類できる．

単一補完法は，被験者毎に最後に観察された値を補完する方法(Last Observation Carried Forward)やベースライン値で補完する方法(Baseline Observation Carried Forward)などが代表的な例である．

多重補完法は，欠測のあるデータについて複数回補完を行い，完全な（欠測値のない）データセットを複数作成し，それらに対して任意の統計手法を適用して，得られた複数の結果を一つに統計的手法を用いて，

統合する方法である．

本論文ではこの多重補完法について以降紹介していく．

6, 多重補完法とは

Rubin, DB(1987)[5]によって提案された方法であり，Multiple imputation(MI)法とよばれる．

MI法の手法については，前項でも述べた通り，欠測値を含むデータに対して，

1. 複数回の補完を行い，
2. 補完後のそれぞれの完全データに対して解析を行い，
3. その結果を1つの最終結果に統合する

過程がとられる(図2参照).

補完の方法には様々な種類があり，各補完方法を用いて複数回の補完を行うことで，欠測値の補完に対しての不確実性を考慮することができる．欠測値を補完するための統計モデルのことを補完モデルとよび，多重補完された完全データを用いて解析するための統計モデルを解析モデルとよぶ.

補完された完全データに対する解析モデルにはAnalysis of variance(ANCOVA)やMixed-effects models for repeated measure(MMRM)が利用されることが多い[6].

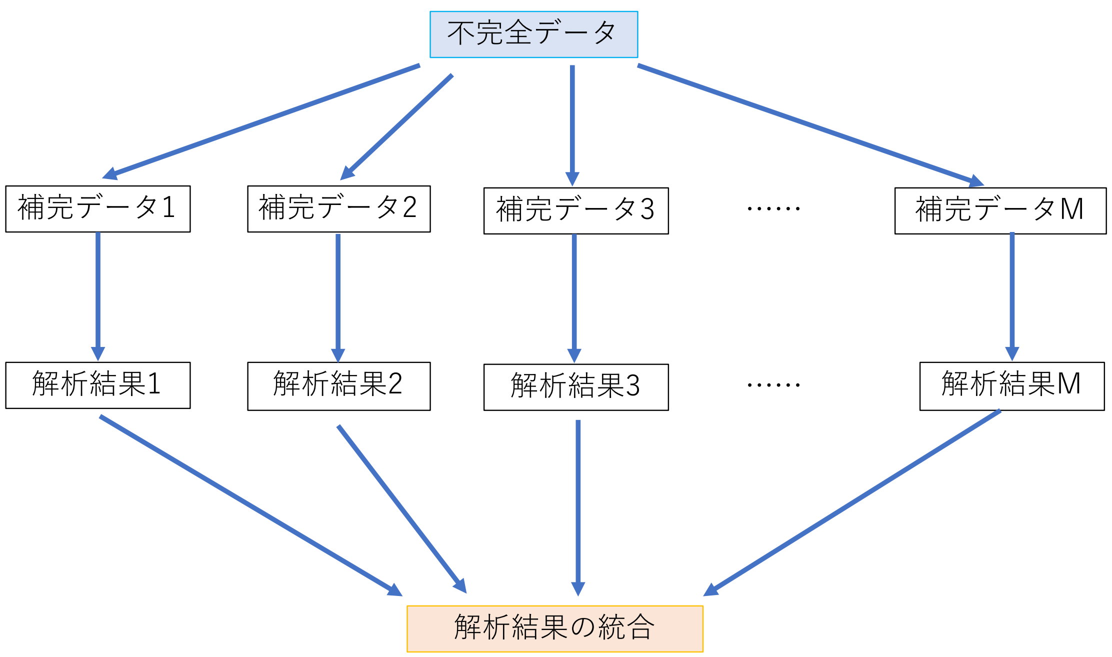


図 2　多重補完(MI)法の手順の図

SASにおいてはMIプロシジャを用いて欠測値の補完を行い，MIXEDプロシジャ等で補完後のそれぞれの完全データに対して解析を行った後，MIANALYZEプロシジャで得られた複数の解析結果を一つに統合するこ

とで，多重補完法を用いた解析の実装が可能となる．次項では欠測値補完のためのMIプロシジャの基本的な操作法について紹介していく．

7, MIプロシジャについて

MIプロシジャは欠測データの欠測値に対して補完を実施する．欠測パターンが単調，非単調問わず補完可能であり，補完モデル内の変数がカテゴリカルでも連続値でも補完可能である(表1参照).

表 1　欠測を含む連続量経時測定データに対する MI プロシジャによる補完方法(駒嵜弘，藤原正和(2016)より引用(一部改変))

欠測パターン	補完モデル内の変数	補完方法
単調のみ	連続値	Sequential regression imputation (Monotone) 欠測値に対して, その時点までに得られている観測値から予測するための解析モデルを構築して補完する.
非単調も可	連続値	Joint modeling approach (マルコフ連鎖モンテカルロ(Markov chain Monte Carlo (MCMC))) 観測値が与えられたもとで, 多変量分布から計算される欠測値に対する条件付分布から補完値を生成する.
	カテゴリカル, 連続値を含む場合でも可	完全条件付き指定 (Full conditional specification (FCS)) 欠測値のある多変量データの補完を欠測値のある変数ごとに実施する. 各々の欠測値のある変数に対して, 補完モデルを構築し, それぞれの変数に対して補完値を繰り返し作成する.

表2より，MIプロシジャを利用すれば非単調な欠測パターンでも補完が可能である．また，MCMC法で非単調な欠測値を補完して単調な欠測パターンにした後に，Monotone法を用いて単調な欠測パターンとして補完する，という手順も可能である[7].

本論文では欠測メカニズムをMNARに限定し，MIプロシジャの基本操作方法について以下で紹介していく．

8, MIプロシジャの基本操作方法

MIプロシジャの基本構文については以下に記載する．MIプロシジャのステートメント，オプションについては数が膨大であるため，本論文では基本的なステートメント，オプションの紹介に留め，その他のステートメント，オプションについての紹介は割愛する．その他のステートメント，オプションについての詳細は，SAS/STAT® 15.1 User’s Guide (2019) [8]を参照のこと．

```
PROC MI DATA =[データセット名]　 SEED=[数字]　NIMPUTE= [数字]　OUT=[出力データセット名];
CLASS TRT;
MONOTONE method (imputed <= effects> / options) ;
MCMC (option);
FCS NBITER = [数字];
MNAR MODEL (Y1/ MODELOBS =CCMV(K=k))
          MODEL(Y2 / MODELOBS = NCMV(K=k))
          MODEL(Y3/ MODELOBS = (TRT="t"))
          MODEL(Y4/ MODEL option)
;
VAR Y0 Y1 Y2 Y3 Y4;
RUN;

【記号説明】
Y0: ベースライン時の観測値（必ず測定）
Y1,Y2,Y3, Y4: ベースライン後の観測値
TRT: 投与群(t =1, 2)
```

表2　MI プロシジャの基本的なステートメントとオプション(一部)

ステートメント	オプション	値	説明
	SEED	数字	乱数を利用するアルゴリズムを用いた解析で結果を再現するために指定する．同じSEED値を指定することで同じ結果を得ることが可能になる．
	NINPUTE	数字	欠測値の補完回数を指定．デフォルトでは25．推定効率の観点では3~5回でも十分とされているが，例えば，欠測値がデータ全体の20％占める場合には20回以上，少なくとも欠測確率以上の補完回数が必要だという指摘もある[9].
MONOTONE			Sequential regression imputationで欠測値の補完を行う．後述のMETHODオプションで，欠測値を予測するための解析モデルを指定する必要がある．
MONOTONE	(METHOD)	[METHOD名]/([補完対象の変	MONOTONEステートメントにおいて，欠測値を予測するための解析モデルを指定する．

		数名))	REG(単調回帰法), LOGISTIC(ロジスティック回帰法)など．詳しくはSAS/STAT® 15.1 User's Guide (2019) [8]を参照のこと．
MCMC			マルコフ連鎖モンテカルロ(MCMC)法で欠測値の補完を行う． MNARメカニズム仮定の下では使用できないが，非単調な欠測データの場合，MCMCステートメントを利用し，単調な欠測データへと補完した後，MONOTONEステートメントを使用し，MNARメカニズムの欠測データの補完を実施する方法もある[7].
MCMC	IMPUTE	FULL/ MONOTONE	補完する欠測値を指定する．すべての欠測値を補完する場合はFULLを指定し, 非単調パターンの欠測データセットを単調パターンのデータセットにするための補完を行う場合はMONOTONEを指定する．MONOTONEが指定された場合，VAR変数に記載された変数の順番に欠測値補完が行われ，単調パターンの欠測データセットを作成する． デフォルトはFULL.
FCS			完全条件付き指定(FCS)法で欠測値の補完を行う． 各々の欠測値のある変数に対して, 補完モデルを構築し, それぞれの変数内の欠測値に対して補完値を繰り返して作成し，補完を行うため，変数内での補完回数をNBITERオプションで指定する必要がある．
FCS	NBITER	数字	変数内での欠測値に対しての補完回数を指定する． デフォルトは20.
MNAR		-	データの欠測メカニズムがMNARであると仮定して欠損値を代入する．MNARステートメントは, FCSステートメントまたはMONOTONEステートメントのいずれかを指定した場合にのみ適用される．
MNAR	MODEL	([補完を行う変数名]/補完方法)	補完する変数及び補完する変数の補完方法を指定する． MONOTONEステートメントまたはFCSステートメントのいずれかのステートメントが指定されていないと利用

			不可である．補完方法にはComplete Case Missing Value(CCMV), Neighboring Case Missing Value(NCMV), (obs-variable=character-list)の3種類が指定可能である*1*2.
VAR			解析に使用する変数を指定する．VARステートメントを省略する場合，データセット内に含まれるすべての連続変数が指定されることと同義になる．
CLASS			VARステートメントで指定された変数のうち，分類変数を指定する．MCMCステートメントでは指定できない．

*1：：FCSステートメント指定の場合は(obs-variable=character-list)の1種類のみ使用が可能，MONOTONEステートメント指定の場合はどの補完方法でも使用可能となっている．MODELオプションは複数行入力でき，複数の補完変数を指定可能であるが, 同一補完変数を複数のMODELオプションに設定することは不可能である．また，欠測があるが補完変数に指定されていない変数に関しては自動的にMARメカニズムとして補完される．

*2：MODEL ([補完を行う変数名] / model-options)オプションのmodel optionについて

MNARステートメントにおけるMODELオプション内のmodel optionには，①Complete Case Missing Value，②Neighboring Case Missing Value，③ The subset of observations from which the imputation modelsの3種類の補完方法が選択できる．以下ではその3種類の簡単な説明を記載する(渡邊(2016)[4]から一部引用(一部改変)).

① MODELOBS = CCMV <K=k> : Complete Case Missing Value

完全データから欠測データを補完する． “K=”で，任意の時点まで観測されているデータを欠測データ推測のために指定することができる．デフォルトはK=1で，最終時点まで観測されているデータ(完全データ)のみ，欠測データの推測に使用する．

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 1 CCMV における k=1 のときの欠測データ推測の方法

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 2 CCMV における k=2 のときの欠測データ推測の方法

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 3 CCMV における k=3 のときの欠測データ推測の方法

② MODELOBS = NCMV <K=k> : Neighboring Case Missing Value

注目する欠測データのパターンに最も近い欠測パターンのデータから，欠測データを補完する．“K=”で，k番目に近い欠測パターンを用いて欠測データの補完を行う．デフォルトはK=1で，最も近い欠測パターンから，欠測データを補完を実施する．

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 4 NCMV における k=1 のときの欠測データ推測の方法

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 5　NCMV における k=2 のときの欠測データの推測方法

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

図 6　NCMV における k=3 のときの欠測データの推測方法

③　MODELOBS = (obs-variable=”character-list”) : The subset of observations from which the imputation models.

特定のCLASS変数の値を有するデータで補完を行う場合に使用する. 例えば, 治療を表す変数trtは1(実薬), 2(プラセボ)の2種類の値が格納されているとし，図7のようにtrt=1,2それぞれで欠測パターンがA~Dまで発生していたとする．このときMODELオプション内でtrt=”2”と指定することで，trt=2(プラセボ)でのデータで補完モデルを作成し，その作成された補完モデルで「両群の」欠測データを補完する方法をとっている.

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

trt1のデータ

	Y0	Y1	Y2	Y3
パターンA	○	○	○	○
パターンB	○	○	○	×
パターンC	○	○	×	×
パターンD	○	×	×	×

trt2のデータ

図 7 (obs-variable=character-list)のときの欠測データ補完方法

9, 補完された欠測データの解析

MIプロシジャによって補完された欠測のある経時的測定データは，完全データとしてANCOVAやMMRMなどの手法を用いてそれぞれ解析される．補完回数分生じたデータセットに対して，例えばBYステートメントなどで並列に同処理の解析を実施するイメージになるが本論文では具体的SASコードまでは言及しない．

10, MIANALYZEプロシジャの基本操作

MIプロシジャによって多重補完された経時的測定データのそれぞれの完全データの解析結果の統合を，MIANALYZEプロシジャによって実施する．

MIANALYZEプロシジャの基本構文は以下になる．

```
PROC MIANALYZE PARMS = [多重補完されたデータの解析結果の出力データセット]
MODELEFFECTS [対象推定値の変数];
ODS OUTPUT ParametersEstimates = [解析結果]
RUN;
```

11, 多重補完法のSASでの実装例

これまで，欠測メカニズムにMNARが仮定される経時的測定データ欠測データについてのSASにおける多重補完法として，MIプロシジャとMIANALYZEプロシジャの基本的な構文を紹介してきた．

本項では，シミュレーションデータを用いて実際にMIプロシジャとMIANALYZEプロシジャの挙動と出力結果を確認する．

① シミュレーションデータの作成

経時的測定データの完全データと，その完全データに対しさらにMNARを仮定するシミュレーションデータを発生させる．シミュレーションデータにMIプロシジャを用いて補完データを取得し，補完データに対しMMRMで解析を行った後，得られた解析結果をMIANALYZEプロシジャで統合する．

シミュレーションデータ発生方法については横山(2016)[10]を参考にした．

シミュレーションデータは，大うつ病性障害患者を対象にベースラインを含んだ4時点でハミルトンうつ病評価尺度(HAM-D)を経時的に測定し，実薬群とプラセボ群の間の最終時点でのHAM-Dスコアの2群比較を想定する．各時点の測定値の平均と標準偏差については表3に従うものとし，MNARの欠測メカニズムについては以下を仮定した．また，欠測パターンについては，非単調な欠測を仮定した．

$$logit(p_t) = -9.5 + 0.2 * y_{t-1} + 0.2 * y_t$$

・t *：時点*$(t = 1, 2, 3)$

・y_t *：時点*t*の測定値*

・p_t *：時点*$t-1$*で観察された症例が時点*t*で欠測する確率*

以上の条件のシミュレーションで発生した完全データと非単調な欠測データの各時点の要約統計量についてはそれぞれ表4と表5の結果になった．

表 3 各時点の測定値の平均値

測定値の平均(標準偏差)	ベースライン(y0)	時点1 (y1)	時点2 (y2)	時点3 (y3)
実薬群	20.0 (4.0)	15.0 (5.0)	12.5 (6.0)	11.0 (7.0)
プラセボ群	20.0 (4.0)	16.0 (5.0)	14.0 (6.0)	13.0 (7.0)

表 4 シミュレーションにおける完全データの要約統計量

平均値(標準偏差)	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
実薬群	20.03 (4.38)	15.52 (5.04)	12.55 (6.30)	11.70 (7.35)
プラセボ群	19.85 (4.35)	15.92 (4.81)	13.21 (5.93)	13.52 (6.95)

表 5 シミュレーションにおける非単調な欠測データの要約統計量

平均値(標準偏差)	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
実薬群	20.03 (4.38)	14.70 (4.70)	12.10 (6.05)	10.98 (6.81)
プラセボ群	19.85 (4.35)	14.95 (4.45)	12.85 (5.78)	13.29 (6.84)

② シミュレーションデータの欠測値補完

前項で作成したシミュレーションデータに対してMIプロシジャを用いて欠測値補完を行った.

非単調欠測を仮定するため，MCMCで単調欠測データになるように補完したのちMONOTONE(単調回帰法)で欠測値補完を実施した. MONOTONE(単調回帰法)での欠測値補完においては，CCMVの補完法を実施した. なお，MIプロシジャのMNARステートメントのMODELオプションで指定する際には，CCMVにデフォルト(K=1)の条件を指定した.

表6に，本論文で実施する，欠測なしの条件を含めた2種類のパターンの欠測値補完法をまとめた.

表 6　本論文で実施する欠測値補完法（完全データの場合を含む）

補完パターン	欠測パターン	ステートメント	Method オプション	MNARステートメントの MODELオプション
A	欠測なし	-	-	-
B	非単調欠測	MCMC + MONOTONE	REG	CCMV(K=1)

以下の図8~14において，MIプロシジャを用いて，MCMCで単調欠測データになるように補完したのちMONOTONE(単調回帰法)で欠測値補完を行った出力結果を示す(補完パターンB).

図8~10はMCMCで単調欠測データになるように補完を行った際の出力結果である. 図8で補完の際の補完モデル，補完回数，シード値等の情報が表示される. 図9では，補完される非単調欠測データの欠測パターン一覧と，欠測パターンと時点毎の平均値一覧が表示されている. 図10では補完後の各時点の平均値と時点同士の分散共分散構造が表示されている.

モデルの情報	
データセット	WORK.D3
方法	Monotone-data MCMC
複数の補完連鎖	Single Chain
MCMC の初期推定	EM Posterior Mode
開始	Starting Value
事前	Jeffreys
補完数	30
Burn-in 反復回数	200
反復回数	100
乱数生成のシード	123

図 8 欠測値補完パターン B の，MCMC で単調欠測データに補完した際のモデルの情報

EM (事後最頻値) の推定

TYPE	_NAME_	y0	y1	y2	y3
MEAN		19.938849	14.823712	12.481181	12.107729
COV	y0	18.417265	0.127631	0.915181	-1.609442
COV	y1	0.127631	20.131306	1.031523	-0.243939
COV	y2	0.915181	1.031523	33.852226	-3.791892
COV	y3	-1.609442	-0.243939	-3.791892	46.271394

図 10　欠測値補完パターン B の, MCMC で単調欠測データに補完された補完データの，平均値と分散共分散構造

欠損値データのパターン										
							グループ平均			
グループ	y0	y1	y2	y3	度数	パーセント	y0	y1	y2	y3
1	X	X	X	X	151	75.50	19.302805	14.739621	12.059789	12.210181
2	X	X	X	O	7	3.50	18.640626	15.488177	21.600568	.
3	X	X	.	X	5	2.50	21.026025	14.956016	.	11.005809
4	X	X	O	O	3	1.50	19.572589	17.119254	.	.
5	X	.	X	X	31	15.50	23.368946	.	12.292389	12.240364
6	X	.	X	O	1	0.50	24.159060	.	16.982212	.
7	X	.	.	X	2	1.00	15.058822	.	.	8.623926

図 9　欠測値補完パターン B の，MCMC で単調欠測データに補完される非単調欠測データの欠測パターン一覧と，欠測パターンと時点毎の平均値一覧

図9~14はMCMCで補完された単調欠測データをMONOTONEで完全データに補完を行った際の出力結果である．図8で補完の際の補完モデル，補完回数，シード値等の情報が表示される．図12では，各変数の補完方法(本論文ではCCMV)の情報が示されている．図13では，補完された非単調欠測データの変数の欠測パターンと時点毎の平均値一覧が表示されている．図14では，補完される単調欠測データの欠測パターン一覧と，欠測パターンと時点毎の平均値一覧が表示されている．

モデルの情報	
データセット	WORK.D4_MCMC
方法	Monotone
補完数	30
乱数生成のシード	123

単調モデルの仕様	
手法	補完変数
Regression	y0 y1 y2 y3

図 11　MONOTONE で単調欠測データを補完した際のモデルの情報

MNAR 前提条件の下での補完モデルに使用されるオブザベーション	
補完変数	オブザベーション
y0	Complete Cases
y1	Complete Cases
y2	Complete Cases
y3	Complete Cases

図 12　MONOTONE で単調欠測データを補完した際の各変数の補完方法の情報

欠損値データのパターン											
								グループ平均			
グループ	trt	y0	y1	y2	y3	度数	パーセント	y0	y1	y2	y3
1	X	X	X	X	X	5670	94.50	19.970416	14.766409	12.111837	12.145320
2	X	X	X	X	.	240	4.00	19.330430	15.319167	21.023273	.
3	X	X	X	.	.	90	1.50	19.572589	17.119254	.	.

図 14　MONOTONE で補完された補完データの欠測パターン一覧と，欠測パターンと時点毎の平均値一覧

③ 解析

完全データと欠測値補完データに対してMMRMを実施する．

興味の対象は，最終時点における実薬群とプラセボ群のHAM-Dのスコア差であるため，ベースラインから最終時点までのHAM-Dスコアの変化量で比較を実施する．解析にはMIXEDプロシジャを利用し，有意水準は0.05とする．

完全データでも欠測値補完データでも，MIXEDプロシジャ実行のため，(i)各症例の各時点と補完値を縦1列に並べ，その後に(ii)ベースラインからの変化量の計算及び測定時点の変数を作成．最後に(iii)MIXEDプロシジャでMMRMでの解析を実行する．(i)～(ii)については以下に基本的な構文を記載する((iii)については割愛).

MMRMの実行の際，MIXEDプロシジャのTYPEオプションには「TYPE=UN(無構造)」を指定した．

```
(i)各症例の各時点と補完値を縦 1 列に並べる
PROC SORT DATA=[データセット名]; BY [ソート変数名]; RUN;
PROC TRANSPOSE DATA =[データセット名] OUT=[出力データセット名 1] PREFIX=VAL;
BY [ソート変数名];
VAR Y0 Y1 Y2 Y3;
RUN;

(ii)ベースラインからの変化量の計算及び測定時点の変数を作成
DATA [出力データセット名 2];
SET [出力データセット名 1]; BY [ソート変数名];
TIME = INPUT(SUBSTR(_NAME_, 2), BEST.);
DIFF = VAL1 – VAL0;
RUN;
```

④ 解析結果の統合

前項で得られた各補完データの解析結果をMIANALYZEプロシジャを用いて統合した．これと③により，完全データと欠測値補完データの解析結果を得ることができた．解析結果を表7にまとめる．

表 7 完全データ，欠測値補完データの解析結果

補完パターン	群間差の推定値	群間差の標準誤差	95 % Cl (下限，上限)	p値
欠測なし (A)	-1.8338	1.0144	-3.8342, 0.1667	0.0722
MCMC+MONOTONE での補完 (B)	-2.3683	0.9893	-2.830, -1.940	0.0167

完全データと比較し，欠測データにおける多重補完の群間差はより小さく推定された．しかしながら，欠測値補完は完全データの結果に近づけることが目標ではなく，推定値に欠測によるバイアスが入りにくくすることが目標であるため，一概に今回の多重補完の精度が悪いとはいえない．

本論文では多重補完の手法にMCMC+MONOTONEのみを採用したが，FCSを採用した場合やその他多重補完法との結果とも比較する必要があると考えられる．

12，結論

欠測メカニズムがMNARの場合の欠測値補完法である多重補完(MI)法についての簡潔な紹介を行った．欠測メカニズムがMNARの場合の欠測データについては取り扱いや解析方法に議論が多く交わされる中で，今後さらにMIプロシジャの使用頻度が高くなるのではないかと考えられる．

多重補完は種類が多く，より適切な手法を採用するためには何に着目するべきかについて今後はより深めていきたい．

参考文献

[1] Roderick J. A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, 2002

[2] 日本製薬工業協会, 欠測のある連続量経時データに対する統計手法について ver 2.0, 2016

[3] 野間久志, 欠測データの統計科学：基礎理論と実践的な方法論, 2017

https://www.ism.ac.jp/~noma/2017-01-13%20Lecture%20in%20ISM%20for%20Missing%20Data.pdf

[4] 渡邊大丞, MIプロシジャで実行可能なPattern Mixture ModelとMultiple Imputationに基づく解析, 2016

https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2016-a-02-06.pdf

[5] Rubin, D. B., Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons., 1987

[6] 藤原 正和, 高橋文博, 欠測のあるデータに対する解析手法の基礎 ～(2) 主解析の検討～, 2015

https://www.biometrics.gr.jp/news/all/seminar_2015-2.pdf

[7] 駒嵜弘，藤原正和欠測を含む順序カテゴリカル経時データの解析 -MIプロシジャの有用性-, 2016

https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2016-c-06.pdf

[8] SAS/STAT® 15.1 User's Guide The MI Procedure, 2019

[9] White, I. R., Royston, P., & Wood, A. M., et al., Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 2001, 30(4), 377‐399.

[10] 横山雄一，大浦智紀，土井正明, MIプロシジャで実行可能なPattern Mixture ModelとMultiple Imputationに基づく解析, 2016

https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2016-a-02-06.pdf

付録：シミュレーションデータ

```
data d0;
        /*乱数固定*/
        call streaminit(1);
        do i = 1 to 100;
                y0 = rand("normal", 20, 4);
                y1 = rand("normal", 15, 5);
                y2 = rand("normal", 12.5, 6);
                y3 = rand("normal", 11, 7);
                trt=1;
                output;
        end;
        do i = 1 to 100;
                y0 = rand("normal", 20, 4);
                y1 = rand("normal", 16, 5);
                y2 = rand("normal", 14, 6);
                y3 = rand("normal", 13, 7);
                trt=2;
                output;
        end;
run;

data d1;
format id;
do i= 1 to 100;
        id= i ; trt=1; id = 1000+i; uni =ranuni(4696); output;
        id= i ; trt=2; id = 2000+i; uni =ranuni(5096); output;
end;
run;

proc sort data=d0; by i trt; run;
proc sort data=d1; by i trt; run;
data d2;
merge d0 d1; by i trt;
drop i;
run;
```

```
data d3;
set d2;
call streaminit(123);
array y{4} y0-y3;
array p{3} p1-p3;
array m{3} m1-m3;
do t=1 to 3;
        p{t}=1 / ( 1 + exp( - (-9.5 + 0.2*y{t} + 0.2*y{t+1})));
        m{t}=rand('bernoulli',p{t});
end;
if m1=1 then do;    y1=.; end;
if m2=1 then do; y2=.;   end;
if m3=1 then do; y3=.; end;
keep id trt y0-y3;
run;
```

MMRM入門

○飯田　絢子
（イーピーエス株式会社）

First Steps in mixed effect models for repeated measures (MMRM)

Ayako Iida
EPS Corporation

要旨

mixed effect models for repeated measures (MMRM)について紹介する．MMRM は，直訳すると反復測定データに対する混合効果モデルであり，主に以下の 3 つの特徴がある．本稿では，これら 3 つの特徴について詳細を述べると共に，一般線形モデルの問題点と，どのように線形混合モデルにて問題を解消しているかについて触れる．また，線形混合モデルの中での MMRM の位置づけについて触れる．最後に SAS 実装方法及び注意点を紹介する．

MMRM における 3 つの特徴

(1)　線形混合モデルの一種である

(2)　反復測定データにおける，同一個体内のデータの独立性を仮定せずに柔軟化できる

(3)　欠測データがあっても解析できるが，欠測メカニズムによっては結果にバイアスが生じ得る

キーワード：固定効果，変量効果，MMRM，線形回帰，分散共分散構造，欠測

１．　背景

医薬品開発の臨床試験において，治療効果の評価は，経時的に測定されたアウトカムに基づいて行われることが多い．経時的データは，同一個体において複数の評価時点で反復測定されるデータである為，通常，同一個体におけるアウトカムの独立性が成り立たず，相関構造を考慮する必要がある．また，反復測定される場合は特に，欠測が生じ得る為，欠測データの取り扱いには注意を要する．一般線形モデルでは，欠測データが存在する場合，1 時点でも欠測をもつ症例は，症例ごと解析から除外される問題がある．そこで，欠測データを持つ被験者の情報を利用できるよう，単一の補完手法として，欠測直前に測定された値を補完に用いる Last Observation Carried Forward (LOCF) 手法が使用されてきた. しかし, 非常に強い仮定であることから, 多くの議論がされている[1]．また，LOCF といった単一補完法は，前提となっている仮定が科学的に妥当でない限り，主要な方法として使うべきではないと言われている[2].

欠測を伴う経時測定データの解析においてこれらの問題を解決する為，線形混合モデルは，ひとつの標準的な方法として広く用いられてきた[3]．中でも MMRM は，被験者単位の変量効果を被験者内の誤差相関構造の一部としてパラメータ化する点が特徴である[1]．これにより，被験者内の誤差相関の強さ及び欠測データをもつ被験者の観測データが平均からどれだけ乖離しているかによって欠測値の調整を行う[1]為，欠測の補完データを明示せずに解析ができ，LOCF の問題も解決できる．

MMRM はその実践的有用性により急速に普及している[2]．本稿においては，近年普及してきている MMRM

に関し入門編として，上記要旨にて記載した 3 点について紹介すると共に，SAS 実行例を交えながら解説していく．

２． MMRM の 3 つの特徴について

(1) 線形混合モデルの一種である

本稿では，ベクトル及び行列の文字を太文字で記載する．線形混合モデルの解説の前に，比較の為，一般線形モデルから述べる．

一般線形モデルの式は，$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$で表される．p 次元の固定効果変数で構成されるデータが n 個あるとすると，$\mathbf{Y}$は観測された従属変数を表す n 行 1 列ベクトルである．$\mathbf{X}$は n 行 p 列 の固定効果の計画行列である．$\boldsymbol{\beta}$は p 行 1 列の未知の固定効果ベクトルである．$\boldsymbol{\varepsilon}$は n 行 1 列の観測されない誤差のベクトルである．$\boldsymbol{\varepsilon}$の各要素は，無相関で，平均 0，分散σ^2の正規分布である（以降，行列の各要素の分散が全てσ^2で対角要素が全て 0 である，等分散σ^2で無相関の要素を持つ行列をσ^2I とする）．

線形混合モデルの式は，$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$で表される．$\mathbf{Z}\boldsymbol{\gamma}$項以外は一般線形モデルの説明と同様である．

例えば，一般線形モデルの例で用いた，観測していない何かしらの要素$\boldsymbol{\varepsilon}$の中身が，互いに相関し合っているような場合，一般線形モデルで述べた，$\boldsymbol{\varepsilon}$の各要素は，無相関で，平均 0，分散σ^2の正規分布であるという前提が使えなくなる．そうした場合，変量効果として$\boldsymbol{Z\gamma}$項を加える概念が線形混合モデルである．

q 個次元の変量効果変数で構成されるデータが n 個ある場合，$\boldsymbol{Z}$は n 行 q 列の変量効果の計画行列であり，$\boldsymbol{\gamma}$は n 行 1 列の未知の変量効果ベクトルである．変量効果ベクトル $\boldsymbol{\gamma}$は，確率変数として平均 0，分散共分散行列$\mathbf{G}$を持つ．分散共分散行列$\mathbf{G}$には，構造のバリエーションを指定することができる．また，(2)でも詳細を述べるが，誤差項$\boldsymbol{\varepsilon}$にも分散共分散行列$\mathbf{R}$を設定できる．

ここで，$\boldsymbol{\gamma}$と$\boldsymbol{\varepsilon}$は，互いに独立で平均は 0，分散はそれぞれ分散共分散行列$\mathbf{G}$，$\mathbf{R}$に従うことから，

期待値$E\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$, 分散$V\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix}=\begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$であり，$\mathrm{V}[\mathbf{Y}]=\mathbf{ZGZ'}+\mathbf{R}$と示される．なお，$\boldsymbol{\gamma}$と$\boldsymbol{\varepsilon}$は正規分布に従う為，

$\gamma_{\mathrm{i}}\ \sim\mathrm{N}\,(0,\mathbf{G})$，$\varepsilon_{\mathrm{i}}\ \sim\mathrm{N}\,(0,\mathbf{R})$ である．

MMRMは，分散共分散行列 $\mathbf{V_i}$ を直接的にモデリングすることが考えられる．被験者単位の変量効果を被験者内の誤差相関構造の一部としてパラメータ化することが，MMRM と他の一般線形モデルを区別する特徴である[1].

$\mathbf{Y_i}\ \sim\mathrm{N}\,(\mathbf{X_i}\boldsymbol{\beta},\mathrm{V_i})$，

$\mathbf{V_i}=\mathbf{Z_i}\mathbf{G}\mathbf{Z_i}'+\mathbf{R}$

(2) 反復測定データにおける，同一個体内のデータの独立性を仮定せずに柔軟化できる

従来の一般線形モデルにおいて誤差は等分散σ^2で無相関の要素を持つことを前提としていたが，MMRM では相関構造等を指定できるようになった（$\boldsymbol{\varepsilon}$は平均がＥ[$\boldsymbol{\varepsilon}$]=0 で V[$\boldsymbol{\varepsilon}$] の分散共分散構造$\mathbf{R}$が設定できる）．このことで，反復測定データのような，独立でないデータも相関等を考慮した上で解析できるようになった．

(3) 欠測データがあっても解析できるが，欠測メカニズムによっては結果にバイアスが生じ得る

MMRM は欠測を持つデータがあっても解析できる．従来の一般線形モデルの glm プロシジャでは，反復測定データで 1 時点でも欠測をもつ症例は，症例ごと解析から除外されていたが，mixed プロシジャを使った MMRM では含められる．一般線形モデルでデータの欠測は，慎重に取り扱わなければ誤った結果につながる

恐れがある．LOCF といった単一補完法は，前提となっている仮定が科学的に妥当でない限り，主要な方法として使うべきではないとされている[2]．一方で MMRM において欠測データを持つ被験者のデータは，解析に含められる為，臨床試験においては主要な解析に採用されることも多い[2].

欠測の種類は Missing Completely At Random (MCAR)，Missing At Random (MAR)，Missing Not At Random (MNAR) の 3 種類に大別される．MCAR は完全にランダムに発生しており，観測値とも欠測値とも独立である．MAR は，欠測するかどうかの確率が，観測値には依存するが欠測値には依存しない．MNAR は欠測するかどうかの確率が，観測値にも欠測値にも依存する[4].

MMRM では欠測データがあっても解析できると述べたが，試験デザインや計画上，欠測のメカニズムに MNAR の想定が強く支持される場合は，バイアスを生じ得る．MMRM を MAR または MCAR 以外で使用する場合は，主解析ではなく感度分析として利用するのがよいとされる[2]

３． 各種線形モデルの SAS 実装方法と結果の見方

3.1 線形混合モデル (MMRM 以外の線形混合モデル)

3.1.1 線形混合モデルを扱う為の SAS 基本構文 (MMRM 以外の線形混合モデル)

線形混合モデル$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$について以下の例を考える．$\mathbf{X}$は固定効果に対応する計画行列，$\boldsymbol{\beta}$は固定効果パラメータベクトル，$\mathbf{Z}$は変量効果に対応する計画ベクトル，$\boldsymbol{\gamma}$は変量効果のパラメータ，$\boldsymbol{\varepsilon}$は誤差ベクトルである．

```
proc mixed data=[dataset];
    class A B Block;
    model Y = A B A*B;
    random Block A*Block;
run;
```

PROC MIXEDステートメント及びMODELステートメントは必須である．

- [dataset] ：使用するデータセット
- proc mixed ステートメントにおけるオプション
  - Method オプション : 推定方法を指定する．「REML」 (restricted maximum likelihood．制限付き最尤推定) または「ML」 (maximum likelihood．最尤推定) を設定できる．デフォルトでは REML である．
- Class ステートメント：分類変数として扱う変数を記載する
- Model ステートメント：モデルに含まれる応答変数と固定効果の関係を特定する
- Model ステートメントにおけるオプション
  - ・DDFM オプション : 自由度の計算方法を指定する．Random ステートメントを設定している際は，デフォルトで CONTAIN が指定されており，他にも様々な指定が可能である．ただし欠測によりデータがアンバランスであるとき，Kenward-Roger (KR) 法の使用が推奨されている[1].
  - ・Solution (S) オプション：固定効果解ベクトルの表示を要求する
- Model ステートメント：固定効果としてモデル行列$\mathbf{X}$に含める変数を記載する

- Random ステートメント：変量効果としてモデル行列$\mathbf{G}$に含める変数を記載する
- Random ステートメントにおけるオプション

  TYPE オプション：変量効果の分散共分散構造$\mathbf{G}$の構造を指定する．分散共分散構造の全成分に異なる値が入る，Unstructured (UN) 構造または非対角成分が 0 の variance components (VC) 構造が指定できる．デフォルトでは VC 構造である．
- 切片：デフォルトでは，すべてのモデルには，切片パラメータμ ($X_i = 0$の時のY_iを集めた平均値$\overline{Y}_{..}$)

  を推定するために，固定効果に対応する計画行列$\mathbf{X}$に 1 の列が自動的に含まれる．対照的に，変量効果に対応する計画ベクトル$\mathbf{Z}$にはデフォルトでは切片が含まれていない．もし変量効果計画ベクトル$\mathbf{Z}$に切片を含めたい場合は，RANDOM ステートメントで， INTERCEPT を変量効果の対象にする必要がある．

3.1.2 主な output について (MMRM 以外の線形混合モデル)

The MIXED Procedure の Examples における Example 83.1 Split-Plot Design を用いた例を示す．例については SAS® 9.4 および SAS® Viya® 3.5 プログラミングドキュメントの The MIXED Procedure における，Example 84.1 Split-Plot Design [5]を用いた．

(https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/statug/statug_mixed_examples01.htm)

分類変数の水準の情報		
分類	水準	値
A	3	1 2 3
B	2	1 2
Block	4	1 2 3 4

次元の数	
共分散パラメータ	3
X の列	12
Z の列	16
サブジェクト	1
サブジェクト毎の最大 OBS	24

オブザベーション数	
読み込んだオブザベーション	24
使用されたオブザベーション	24
使用されなかったオブザベーション	0

分類変数の水準の情報：

データセットにおいて分類変数に指定した A，B，Block の各水準が記載される．A，B，Block の水準は順に 3 水準，2 水準，4 水準である．

次元の数：

共分散パラメータの数は，random ステートメントで指定した，Block，A*Block の 2 変数と，残差の合計 3 次元である．行列$\mathbf{X}$の列数は，切片 1 列， model で指定した，固定効果 A の水準 3 列及び B の水準 2 列，A*B の水準の組合せ数 6 列を全て足して合計 12 次元というように計算する．同様に，$\mathbf{Z}$の列数は，random ステー

トメントにおいて変量効果として指定した，Block の水準 4 列と，A*Block の組合せ 12 列を足して，16 次元となる．

共分散パラメータの推定	
共分散パラメータ	推定値
Block	62.3958
A*Block	15.3819
Residual	9.3611

共分散パラメータの推定結果：
random ステートメントにて変量効果に指定した Block，A*Block，及び残差の推定結果がそれぞれ出力される．

固定効果の Type 3 検定				
効果	分子の自由度	分母の自由度	F 値	Pr > F
A	2	6	4.07	0.0764
B	1	9	19.39	0.0017
A*B	2	9	4.02	0.0566

固定効果の Type 3 検定結果：
model ステートメントにて固定効果に指定した，A，B，A*B の推定結果がそれぞれ出力される．
推定には最尤法または制限付き最尤法が用いられる．帰無仮説は各固定効果のベクトル成分=0 である．

3.2 MMRM

3.2.1 反復測定データに対する混合効果モデル (MMRM) の SAS 基本構文

例については SAS® 9.4 および SAS® Viya® 3.5 プログラミングドキュメントの The MIXED Procedure における，Example 84.2 Repeated Measures [6]を用いた．
(https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/statug/statug_mixed_examples02.htm)

```
proc mixed data=[dataset] method=ml covtest;
    class Person Gender;
    model y = Gender Age Gender*Age / solution;
    repeated / type=un subject=Person r;
run;
```

PROC MIXED ステートメント及び MODEL ステートメントは必須である．各ステートメントとオプションの一例を挙げる．3.1.1 項にて述べたステートメント及びオプションについては割愛する．

- Repeated ステートメント：誤差の分散共分散行列**R**行列の対象変数を指定する．
- Repeated ステートメントにおけるオプション

・TYPE オプション：誤差の分散共分散構造**R**の構造を指定する．UN 構造，VC 構造，Autoregressive (1) (AR1) 構造等，指定できる．デフォルトでは VC 構造である．

・Subject オプション：誤差の分散共分散構造**R**における，繰り返し測定の単位を指定する

・R オプション：Repeated ステートメントで指定した構造において，表示させたい症例ブロックを指定する．例えば，r=1，3，5 とすると，1，3，5 番目の症例における 4×4 成分で構成されるブロックが表示される．なお，r オプションをつけて，以降の数値を指定しない場合は，デフォルトで，1 症例目のブロックが表示される．どの症例も同じブロックを持ち，症例ブロック同士は相関なし，症例ブロック内は，誤差の分散共分散行列**R**の構造として指定した形の相関等の関係性が設定される．

3.2.2　主な output について（MMRM)

分類変数の水準の情報

分類	水準	値
Person	27	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Gender	2	F M

次元の数	
共分散パラメータ	10
X の列	6
Z の列	0
サブジェクト	27
サブジェクト毎の最大 OBS	4

分類変数の水準の情報：

本データは，男女合計 27 人のデータで構成され，即ち分類変数に指定した Person は 27 水準，Gender は男性と女性の 2 水準である．

次元の数：

Class ステートメントに Age を設定していないため，Age は連続量データとしている．

Xの列は，先ほどまでの紹介と同様の計算で，切片 1 列＋Gender の 2 水準＋Age の 1 水準（連続量の為，1 水準と考える）＋Gender*Age の組合せ 2 水準=6 次元となる．

Random ステートメントは宣言していないため，**Z**の列は 0 である．

サブジェクト毎の最大 OBS は，PR データセットでは各症例において 4 時点の年齢(8 歳，10 歳，12 歳，14 歳）を扱っている為，サブジェクト毎に最大 4OBS と考える．このとき，3.1.1 項では順序は気にしていないが 3.2.2 項ではデータの順序に意味があることに注意する．Repeated は経時データのような，順序に時間的意味を持つようなものを扱う点も random ステートメントとの違いの特徴である．また注意点として，例えば欠測データのレコードが発生していない場合は，欠測なしの場合と結果が変わる．

共分散パラメータの推定					
共分散パラメータ	サブジェクト	推定値	標準誤差	Z 値	Pr Z
UN(1,1)	Person	5.1192	1.4169	3.61	0.0002
UN(2,1)	Person	2.4409	0.9835	2.48	0.0131
UN(2,2)	Person	3.9279	1.0824	3.63	0.0001
UN(3,1)	Person	3.6105	1.2767	2.83	0.0047
UN(3,2)	Person	2.7175	1.0740	2.53	0.0114
UN(3,3)	Person	5.9798	1.6279	3.67	0.0001
UN(4,1)	Person	2.5222	1.0649	2.37	0.0179
UN(4,2)	Person	3.0624	1.0135	3.02	0.0025
UN(4,3)	Person	3.8235	1.2508	3.06	0.0022
UN(4,4)	Person	4.6180	1.2573	3.67	0.0001

共分散パラメータの推定：
「repeated / type=un subject=Person r;」の指定が「共分散パラメータ」列に反映されている．type=UN (無構造) に指定していたことから，誤差項の分散共分散の各成分の表示は，「UN（1,1）」というように「UN」が先頭に表示される．同一症例内の分散共分散成分は，時点間の組合せである．即ち，4 時点同士は（1,1），（1,2），（1,3），（1,4），（2,2），（2,3），（2,4），（3,3），（3,4），（4,4）という 10 成分ができる（非対角成分の共分散の値は対角を挟んで対象であり，例えば(1,2) と(2,1) の成分における共分散の値は等しい）．共分散パラメータの推定で表示されるのは，この 10 成分である．
誤差の分散共分散構造は，今回の例では，27 人がそれぞれ 4 時点で測定している為，誤差の分散共分散構造**R**のうち，1 ブロックは 4×4 成分で，誤差の分散共分散構造**R**は，同じものが 27 症例分の 27 ブロックある．各症例同士の誤差は独立しているので， 誤差の分散共分散構造**R**の非対角成分は 0 であるが，同じ症例内では誤差ε_{ij}同士は独立ではない仮定を置いている (例では UN 構造を指定している) 為，症例ブロック内の非対角成分は，0 以外の数値が入り得る．ただし構造指定によっては，非対角成分を 0 ともできる．
成分の見方については，例えば UN (1,1) ， UN (2,2) ， UN (3,3) ， UN (4,4) はそれぞれ 8 歳，10 歳，12 歳，14 歳の時の誤差項の分散を意味し，UN (2,1) は 8 歳の時と 10 歳の時の誤差項の共分散を意味する．

標準誤差：共分散パラメータの漸近的標準誤差．各共分散パラメータに対する尤度の二次微分行列の逆数から計算される．

共分散パラメータ推定値を標準化した Z 値：推定値を標準化した値であり，下記の式で示される．

$$\text{共分散パラメータ推定値を標準化した Z 値} = \left(\frac{\text{推定値} - \text{平均}}{\text{標準誤差}}\right)$$

wald 検定 : 上記より求めた，共分散パラメータ推定値を標準化した Z 値を 2 乗した値が，自由度 1 の χ 二乗分布に従うことを利用している．

$$\text{Wald} = \left(\frac{\text{推定値} - \text{平均}}{\text{標準誤差}}\right)^2$$

検定の帰無仮説：個別パラメータ推定値が 0 である

固定効果の Type 3 検定				
効果	分子の自由度	分母の自由度	F 値	Pr > F
Gender	1	25	1.17	0.2904
Age	1	25	110.54	<.0001
Age*Gender	1	25	7.99	0.0091

Person 1 の推定 R 行列				
行	Col1	Col2	Col3	Col4
1	5.1192	2.4409	3.6105	2.5222
2	2.4409	3.9279	2.7175	3.0624
3	3.6105	2.7175	5.9798	3.8235
4	2.5222	3.0624	3.8235	4.6180

Person 1 の推定**R**行列：

R オプション指定により，指定された症例ブロックが表示される．先ほどの共分散パラメータの推定結果をブロック形式で表示している．

固定効果の解						
効果	Gender	推定値	標準誤差	自由度	t 値	Pr > |t|
Intercept		15.8423	0.9356	25	16.93	<.0001
Gender	F	1.5831	1.4658	25	1.08	0.2904
Gender	M	0	.	.	.	.
Age		0.8268	0.07911	25	10.45	<.0001
Age*Gender	F	-0.3504	0.1239	25	-2.83	0.0091
Age*Gender	M	0	.	.	.	.

固定効果の解：

Solution (S) オプションにより，固定効果の推定値等が表示される．

推定値の計算方法については以下の通りである．

Gender

M (男性) ：Estimate=0→Intercept の値＝15.8423

F (女性) ：Intercept の値＋Estimate の値

＝15.8423+1.5831=17.0654

Age*Gender

Age*M (男性) ：Estimate=0→Age の値＝0.8268

Age*F (女性) ：Age の値＋Estimate の値

＝0.8268+ (-0.3504)＝0.4764

4． MIXED プロシジャの注意点

- PROC MIXED ステートメントにおける EMPRICAL オプション：
  分散共分散 構造に特定の相関構造を仮定しない UN を指定すればパラメータの推定値に偏りは含まれないが，収束の問題等の理由により，より特定した CS 構造などを採用せざるを得ない場合，分散共分散構造を誤って特定すると，欠測データの有無にかかわらずモデルに基づく分散推定量は偏りをもつことが知られている．このような場面では，EMPRICAL オプションを指定し，サンドウィッチ分散推定量を利用する必要がある．サンドウィッチ分散推定量は，平均構造が正しく特定されていれば一致性をもつことが知られている[1].
- MODEL ステートメントにおける DDFM オプション：
  欠測 によりデータがアンバランスであるとき，自由度の計算方法を指定する DDFM オプションでは Kenward-Roger (KR) 法の使用が推奨されている[1].
- REPEATED ステートメントの明記：
  データセッ トの構造上，繰り返しの単位内でデータの順序が自明でない場合 (例えば非単調な欠測で，欠測データのレコードが存在しないデータ構造の場合) は，データの順序の情報をもつ

変数を REPEATED ステートメントで指定することが必要となる．そうでない場合は省略することができる[1].

５． まとめ

従来の一般線形モデルにおいて，以下 3 点の問題について触れた上で，それらの問題を解決する線形混合モデルについて説明した．

(1) 1 点でも欠測データがある場合，症例データごと使用されない

(2) 欠測データを含まない症例のデータも含めるために，LOCF によるデータ補完の下，一般線形モデルを行う手法が用いられてきたが，LOCF による補完は適切でないとする意見がある

(3) 同一個体内の誤差について相関構造が設定できない

MMRM は線形混合モデルでも変量効果を明示せず, 被験者単位の変量効果を被験者内の誤差相関構造の一部としてパラメータ化する点が特徴的で，補完を明示しなくても相関構造や欠測時点までのデータの平均との乖離具合より欠測データを予測し，解析に含めることができる．ただし，MMRM は欠測の種類が MNAR である場合は適さないとされる点，SAS 実装においては欠測値がある場合や UN 以外の構造指定の際に，自由度や分散推定量の指定について注意が必要であることを述べた．

本稿では適合度統計量の解説，参考文献で得た情報についてデータを用いた検証ができなかった為，今後の課題としたい．

６． 参考文献

[1] 日本製薬工業協会，欠測のある連続量経時データに対する 統計手法について Ver.2.0，2016
(https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc0000007qqq-att/statistics01.pdf).

[2] 五所正彦，丸尾和司，経時測定データ解析における mixed-effects models for repeated measures (MMRM)の利用，2017
(https://www.jstage.jst.go.jp/article/jappstat/46/2/46_53/_pdf/-char/ja)

[3] 右京芳文，野間久史，Yoshifumi Ukyo，Hisashi Noma，欠測を伴う経時測定データにおける MMRM（Mixed-Effects Model for Repeated Measures） の並べ替え法に基づく推測手法 Permutation Inference Methods for the MMRM (Mixed-Effects Model for Repeated Measures) in Incomplete Longitudinal Data Analysis，2019
（https://www.jstage.jst.go.jp/article/jjb/40/1/40_15/_pdf/-char/ja）

[4] 高井啓二，星野崇宏，野間久史，調査観察データ解析の実際 1 欠測データの統計科学―医学と社会科学への応用 第 3 版，2018

[5] SAS® 9.4 および SAS® Viya® 3.5 プログラミングドキュメントの The MIXED Procedure Example 84.1 Split-Plot Design
(https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/statug/statug_mixed_examples01.htm)

[6] SAS® 9.4 および SAS® Viya® 3.5 プログラミングドキュメントの The MIXED Procedure Example 84.2 Repeated Measures
(https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/statug/statug_mixed_examples02.htm)

BGLIMMプロシジャおよびMCMCプロシジャによるベイズ流経時測定データ解析

○伊庭　克拓[1]、浅野　豊[1]、松嶋　優貴[1]、毛利　誠[1]

（[1]大塚製薬株式会社 新薬開発本部 バイオメトリックス部 統計解析室）

Bayesian Longitudinal Data Analysis using BGLIMM Procedure and MCMC Procedure

Katsuhiro Iba[1], Yutaka Asano[1], Yuki Matsushima[1], Makoto Mouri[1]

Office of Biostatistics, Department of Biometrics, Headquarters of Clinical Development, Otsuka Pharmaceutical Co., Ltd.

要旨

臨床試験では, 欠測値を伴う経時測定データが得られることが多く, mixed model for repeated measures（MMRM）などの欠測値を考慮した解析が行われている。また，近年，臨床試験でのベイズ流アプローチの利用が注目されている。SAS では，以前より，汎用的なベイズ流の解析を行う MCMC プロシジャが実装されており，SAS/STAT 15.1 からベイズ流の一般化線形混合効果モデルの解析を行う BGLIMM プロシジャが実装されている。本稿では，これらのプロシジャを用いて，抗うつ薬の臨床試験データを事例として，ベイズ流の経時測定データ解析（Bayesian MMRM および回帰モデルに基づく欠測値の補完）を行う方法を紹介する。

キーワード：経時測定データ，ベイズ，BGLIMM，MCMC，Bayesian MMRM

1. はじめに

医薬品や医療機器などの治療法の有効性や安全性を評価する臨床試験では，有効性および安全性に関する評価項目を治療開始前（ベースライン）および治療開始後の複数の時点で評価し，各被験者から経時的にデータを収集することが多い。しかしながら，ほとんどの臨床試験では，治療の中止などで計画されたすべての時点での評価ができないことにより，欠測値が生じる（日本製薬工業協会 2016）。欠測値は，臨床試験の結果に偏りを起こし得る代表的な原因であり（厚生省医薬安全局審査管理課長 1998），発生を防止すると共に，データ解析で適切に対処する必要がある（五所・丸尾 2017）。

主要評価項目が連続変数の臨床試験では，計画された治療終了時点でのベースラインからの変化量の平均値の群間比較を主目的とすることが多く，欠測値を伴う連続変数の経時測定データの解析方法として，mixed model for repeated measures（MMRM）がよく用いられている。MMRM は観測データの尤度に基づい

ており，欠測値の発生メカニズムが missing at random（MAR）のもとで妥当な推測を行うことができる（五所・丸尾 2017）。

臨床試験では頻度流アプローチが用いられることが多いが，近年，既存情報を利用できることや，より柔軟な意思決定を行える可能性があるベイズ流アプローチの利用が注目されている（Food and Drug Administration 2019; Hirakawa et al. 2022）。ベイズ流アプローチは，治療効果などの関心のあるパラメータについて，データを得る前の情報である事前分布を，ベイズの定理を利用してデータの情報である尤度と結合することにより，データを得た後の情報である事後分布に更新する。観測データの尤度に基づいたベイズ流の解析も MAR の仮定のもとで妥当である（Dmitrienko and Koch 2017）。

SAS では，以前より，一般のベイズ流の解析を行うことができる MCMC プロシジャをはじめとして，GENMOD，PHREG，LIFEREG，FMM など，いくつかのプロシジャでベイズ流の解析を行うことができる（SAS Institute Inc. 2023）。さらに，SAS/STAT 15.1（SAS 9.4 TS1M6）から，ベイズ流の一般化線形混合効果モデルの解析を行うことができる BGLIMM プロシジャが実装されている。これらのプロシジャは，いずれもマルコフ連鎖モンテカルロ（MCMC）法によって事後分布からパラメータをサンプリングすることで，ベイズ流の解析を行う。MMRM は線形混合効果モデルの特別な場合であり（五所・丸尾 2017），MIXED プロシジャや GLIMMIX プロシジャで実行できる。BGLIMM プロシジャは，これらのプロシジャと同様に，変量効果をモデルに含めたり，誤差共分散構造をモデル化したりでき，経時測定データ解析に適していると考えられる。そこで，本稿では，BGLIMM プロシジャおよび MCMC プロシジャを用いて，MAR の仮定のもとでのベイズ流の経時測定データ解析を行う方法を紹介する。まず，本稿で用いる事例データの紹介と MIXED プロシジャによる頻度流の MMRM による解析を行った後，BGLIMM プロシジャおよび MCMC プロシジャを用いて，ベイズ流の MMRM（Bayesian MMRM）による解析を行う方法を紹介する。また，別のアプローチとして，MCMC プロシジャを用いて，回帰モデルによって欠測値を補完して解析する方法を紹介する。

2. 本稿で使用する事例データと頻度流 MMRM

本稿では，London School of Hygiene & Tropical Medicine のウェブサイト内の Drug Information Association Scientific Working Group on Estimands and Missing Data のページ（https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data#dia-missing-data）で公開されている事例データ（Chapter15_example.sas7bdat）を用いる。この事例データは，抗うつ薬の有効性を評価するために，患者を 4 つの群（2 用量の被験薬，実対照薬，プラセボ）にランダム化し，うつ病の重症度を評価するハミルトンうつ病評価尺度の 17 項目（HAM-D17）合計スコアを経時的（ベースライン，Week 1, 2, 4, 6, 8）に測定した実際の臨床試験のデータに基づいている（Goldstein et al. 2004）。なお，匿名化のために，事例データでは Week 8 のデータは削除されており，群は 4 群から 2 群に減らされている。データセットの一部を表 1 に示す（投与後のオブザベーションのみ含んでいる）。本稿では，PATIENT（被験者番号），VISIT（時点：4，5，6，7），THERAPY（群：DRUG，PLACEBO），BASVAL（ベースライン時点の HAM-D17 合計スコア），HAMDTL17（HAM-D17 合計スコア），CHANGE（HAM-D17 合計スコアのベースラインからの変化量）の変数を解析に用いる。

以降の節でのベイズ流の解析結果との比較を行うために，事例データを頻度流 MMRM

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i) \tag{1}$$

で解析した結果を示す。ここで，$\mathbf{Y}_i$は被験者iの応答変数ベクトル，$\mathbf{X}_i$は被験者iの固定効果に対する計画行列，$\boldsymbol{\beta}$は固定効果パラメータベクトル，$\mathbf{V}_i$は$\mathbf{Y}_i$の共分散行列である。平均構造$\mathbf{X}_i\boldsymbol{\beta}$には，群，時点，群と時点の交互作用が含まれることが通常であり（五所・丸尾 2017），ベースライン値など有効性に影響する背景因子（および関連する交互作用）も追加されることも多い。共分散行列$\mathbf{V}_i$の構造には unstructured（無構造）を仮定することが一般的である。本稿では，HAM-D17 合計スコアのベースラインからの変化量を応答変数，群，時点，ベースライン値，群と時点の交互作用，ベースライン値と時点の交互作用を固定効果とし，共分散構造は unstructured を仮定した。パラメータ推定は制限付き最尤法（MIXED プロシジャのデフォルト）で行い，固定効果の標準誤差と t 統計量の自由度は Kenward-Roger 法（五所・丸尾 2017）で求めた。事例データを頻度流 MMRM で解析するための SAS のコードを以下に示し，各時点の被験者数と頻度流 MMRM の解析結果（調整済み平均（最小二乗平均）とその群間差）を表 2 に示す。なお，本稿では解析の目的が治療終了時点（Week 6）の平均値の群間比較である状況を想定している。

```
proc mixed data=chapter15_example;
  class THERAPY VISIT PATIENT;
  model CHANGE = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT /ddfm=kenward;
  repeated VISIT /subject=PATIENT type=un;
  lsmeans THERAPY*VISIT /diff cl e;
run;
```

表 1 本稿で使用する事例データ（抜粋）

PATIENT	VISIT	THERAPY	BASVAL	HAMDTL17	CHANGE
1503	4	DRUG	32	21	−11
1503	5	DRUG	32	20	−12
1503	6	DRUG	32	19	−13
1503	7	DRUG	32	17	−15
1507	4	PLACEBO	14	11	−3
1507	5	PLACEBO	14	14	0
1507	6	PLACEBO	14	9	−5
1507	7	PLACEBO	14	5	−9

VISIT: 4 = Week 1, 5 = Week 2, 6 = Week 4, 7 = Week 6

表 2 各時点の被験者数と頻度流 MMRM の解析結果

群	DRUG			PLACEBO			群間差	両側 95%信頼区間	
時点	被検者数	調整済み平均	標準誤差	被検者数	調整済み平均	標準誤差		下限	上限
Week 1	84	−1.61	0.49	88	−1.70	0.47	0.09	−1.26	1.44
Week 2	77	−4.22	0.66	81	−2.82	0.64	−1.40	−3.23	0.42
Week 4	73	−6.37	0.71	76	−4.14	0.70	−2.22	−4.20	−0.25
Week 6	64	−7.62	0.79	65	−4.82	0.78	−2.80	−5.01	−0.60

どちらの群でも Week 1 から Week 6 までに約 20 例が中止していた。また，時点が進むにつれ，両群ともにベースラインからの変化量の調整済み平均値は低下（改善）したが，DRUG 群のベースラインからの変化量の方が大きく，Week 6 での群間差（両側 95%信頼区間）は，−2.80（−5.01，−0.60）であった。

3. BGLIMM プロシジャ

本節では，比較的新しい，一般化線形混合効果モデルに対してベイズ流の推測を行う BGLIMM プロシジャの概要について紹介する。

一般化線形混合効果モデルは，リンク関数を用いることにより，応答変数が指数型分布族に含まれる様々な分布（正規分布，2 項分布，ポアソン分布など）に従うことを仮定できる混合効果モデルである。頻度流アプローチでは，固定効果パラメータは未知の定数であり，周辺尤度を最大化することにより推定される。一方，ベイズ流アプローチでは，固定効果や変量効果を含むモデル内のすべてのパラメータは確率変数であり，MCMC 法により同時事後分布からのサンプリングを行う。BGLIMM プロシジャは，変量効果の分布は正規分布のみであり，固定効果パラメータおよび共分散パラメータの事前分布はオプションで指定できるものに限られているが，広範なモデルに対して，ベイズ流の解析を行うことができる。

BGLIMM プロシジャの基本的な構文を以下に示す。MIXED プロシジャや GLIMMIX プロシジャの構文と類似しており，これらのプロシジャを使用したことがあるユーザーにとっては馴染みやすいと思われる。BGLIMM プロシジャはベイズ流の解析を行うので，事前分布や MCMC の設定をする必要がある。

```
PROC BGLIMM;
  CLASS variables;
  MODEL response = fixed-effects;
  MODEL events / trials = fixed-effects; /* 2項分布データの場合 */
  RANDOM random-effects;
  REPEATED repeated-effect;
  ESTIMATE 'label' estimate-specification;
```

PROC BGLIMM 文の主なオプションを以下に示す。

- NBI　バーンイン（burn-in）の回数を指定する。
- NMC　MCMC のサンプリング回数（バーンインは除く）を指定する。
- SEED　疑似乱数のシードを指定する。
- THIN　間引き（thinning）の回数を指定する。「指定した値」番目ごとのサンプルのみ保持される。
- OUTPOST　MCMC でサンプリングしたパラメータの値を含む出力データセットを指定する。
- DIAG　MCMC の収束診断の出力を指定する。
- PLOTS　トレースプロットなどの MCMC の収束診断プロットの出力を指定する。

MODEL 文は，応答変数と固定効果に含める変数を指定する。主なオプションを以下に示す。

・DIST　応答変数の確率分布を指定する。
・LINK　リンク関数を指定する。
・COEFFPRIOR　固定効果パラメータの事前分布を指定する。デフォルトは無情報フラット事前分布である。無情報の正規分布（平均 0，分散 10^4）も指定できる。また，SAS データセットを読み込ませることで情報のある事前分布を指定することもできる（7 節参照）。
・SCALEPRIOR　スケールパラメータの事前分布を指定する（モデルにスケールパラメータが含まれる場合）。逆ガンマ分布，ガンマ分布，非正則（improper）事前分布を選択できる。

RANDOM 文は，変量効果に含める変数を指定する（変量効果を含めない場合は不要）。RANDOM 文では，切片は INTERCEPT で明示的に指定する必要がある。主なオプションを以下に示す。
・SUBJECT　経時測定の対象（本稿では被験者）などのデータのグループを指定する。
・TYPE　変量効果の共分散行列の構造を指定する。
・COVPRIOR　変量効果の共分散行列の共分散パラメータの事前分布を指定する。逆ウィシャート分布など，6 つの分布を選択できる。TYPE=UN，UN(1)，VC，TOEP(1)の場合のみ有効であり，それ以外の場合はフラット事前分布が用いられる。

REPEATED 文は，経時測定の時点を示す変数と誤差共分散行列の構造を指定する（誤差共分散行列の構造を指定しない場合は不要）。応答変数の分布に正規分布，リンク関数に恒等（identity）リンク関数を指定した場合のみ指定可能である。主なオプションを以下に示す。
・SUBJECT　経時測定の対象（本稿では被験者）などのデータのグループを指定する。
・TYPE　誤差共分散行列の構造を指定する。
・COVPRIOR　誤差共分散の共分散パラメータの事前分布を指定する。逆ガンマ分布と逆ウィシャート分布が利用できる。TYPE=UN，UN(1)，VC，TOEP(1)の場合のみ有効であり，それ以外の場合はフラット事前分布が用いられる。

ESTIMATE 文は，パラメータ（固定効果および変量効果）の線形結合を指定でき，指定したパラメータの線形結合が出力データセットに追加され，その要約統計量などがプロシジャの出力に追加される。なお，BGLIMM プロシジャには LSMEANS 文が存在しないため，調整済み平均を得るためには，ESTIMATE 文でやや煩雑な係数の指定を行う必要がある。

4. BGLIMM プロシジャによる Bayesian MMRM

本節では，BGLIMM プロシジャで Bayesian MMRM による解析を行う方法を紹介する。Bayesian MMRM は，モデルは(1)式の頻度流 MMRM と同じであるが，頻度流 MMRM とは異なり，すべてのパラメータ（固定効果パラメータおよび共分散パラメータ）は確率変数であり，それらに対する事前分布を設定する必要がある。

臨床試験における Bayesian MMRM の事例として，レビー小体型認知症におけるメビダレンの安全性と有効性を評価した第 II 相ランダム化プラセボ対照試験（NCT03305809）では，有効性の主要評価項目について，プラセボに対するメビダレンのエフェクトサイズが 0.2 以上である事後確率を求め，事前規定した閾値

（0.67）を上回るかどうか評価するために Bayesian MMRM が用いられた（Biglan et al. 2022）。また，肺動脈性高血圧症の小児集団におけるタダラフィルの有効性と安全性を評価した第Ⅲ相ランダム化二重盲検プラセボ対照試験（NCT01824290）では，有効性評価の精度を高めるために，事前情報として成人試験のデータを利用した Bayesian MMRM による補足的解析が行われた（Ivy et al. 2021）。

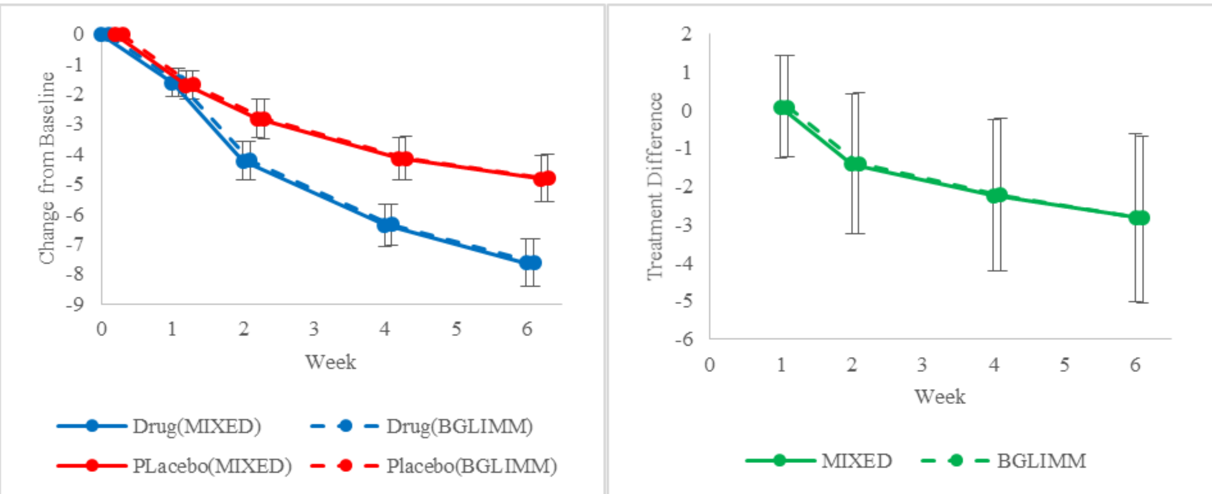
2 節で紹介した事例データに対し，Bayesian MMRM による解析を行った。2 節の頻度流 MMRM と同様に，群，時点，ベースライン値，群と時点の交互作用，ベースライン値と時点の交互作用を固定効果とし，共分散構造は unstructured を仮定した。各固定効果パラメータの事前分布には，平均 0，分散 10^4 の無情報の正規分布$N(0, 10^4)$，共分散パラメータの事前分布に自由度 4，スケールパラメータ 1 の逆ウィシャート分布$IW(4,1)$を用いた。逆ウィシャート分布は，逆ガンマ分布を一般化した分布であり，多変量正規分布の共分散行列に対する共役事前分布であることから，ベイズ流の解析でしばしば用いられる。分布のパラメータである自由度（共分散行列の次元以上）およびスケールパラメータが小さいほど無情報となる（Chen et al. 2016）。なお，TYPE=UN の場合のデフォルトの事前分布は逆ウィシャート分布であるが，BGLIMM プロシジャでの自由度およびスケールパラメータのデフォルト値は，共分散行列の次元+3 である。

事例データを Bayesian MMRM で解析するための BGLIMM プロシジャの SAS のコードを以下に示す。事後分布への収束および事後分布の近似のために必要なバーンインおよびサンプリング回数は，NBI=1000 および NMC=5000 を設定した。DIAG=ALL および PLOTS=ALL の指定で，MCMC のすべての収束診断に関する出力とプロットが得られる（SMOOTH はトレースプロットに平滑化曲線を追加する）。MODEL 文および REPEATED 文は，オプションで事前分布を指定する以外は，MIXED プロシジャとほぼ同じである。3 節でも述べたように，BGLIMM プロシジャには LSMEANS 文がないため，頻度流 MMRM の調整済み平均に対応する推定値（調整済み事後平均値）およびその群間差を得るためには，ESTIMATE 文で MODEL 文の各変数に対応する係数を指定する必要がある。ベースライン値，ベースライン値と時点の交互作用に対する係数 17.857 は，入力データセットの全オブザベーションを用いて算出した BASVAL の平均値である（頻度流 MMRM と Bayesian MMRM の結果を比較するために，MIXED プロシジャの LSMEANS 文の仕様に合わせた）。これらの係数は，MIXED プロシジャの LSMEANS 文の E オプションで確認可能である。なお，交互作用項の係数の順番は，CLASS 文での変数の指定順に依存することに注意が必要である。

```
proc bglimm data=chapter15_example seed=123456 nbi=1000 nmc=5000 thin=1 outpost=POST diag=all
plots(smooth)=all;
  class THERAPY VISIT PATIENT;
  model CHANGE = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT /dist=normal coeffprior=normal;
  repeated VISIT /subject=PATIENT type=un covprior=iwishart(df=4, scale=1);

  estimate 'Week 1 DRUG' INTERCEPT 1 THERAPY 1 0 VISIT 1 0 0 0
                         THERAPY*VISIT 1 0 0 0 0 0 0 0 BASVAL 17.857 BASVAL*VISIT 17.857 0 0 0;
  estimate 'Week 2 DRUG' INTERCEPT 1 THERAPY 1 0 VISIT 0 1 0 0
                         THERAPY*VISIT 0 1 0 0 0 0 0 0 BASVAL 17.857 BASVAL*VISIT 0 17.857 0 0;
  estimate 'Week 4 DRUG' INTERCEPT 1 THERAPY 1 0 VISIT 0 0 1 0
                         THERAPY*VISIT 0 0 1 0 0 0 0 0 BASVAL 17.857 BASVAL*VISIT 0 0 17.857 0;
```

```
  estimate 'Week 6 DRUG' INTERCEPT 1 THERAPY 1 0 VISIT 0 0 0 1
                         THERAPY*VISIT 0 0 0 1 0 0 0 0 BASVAL 17.857 BASVAL*VISIT 0 0 0 17.857;
  estimate 'Week 1 PLCB' INTERCEPT 1 THERAPY 0 1 VISIT 1 0 0 0
                         THERAPY*VISIT 0 0 0 0 1 0 0 0 BASVAL 17.857 BASVAL*VISIT 17.857 0 0 0;
  estimate 'Week 2 PLCB' INTERCEPT 1 THERAPY 0 1 VISIT 0 1 0 0
                         THERAPY*VISIT 0 0 0 0 0 1 0 0 BASVAL 17.857 BASVAL*VISIT 0 17.857 0 0;
  estimate 'Week 4 PLCB' INTERCEPT 1 THERAPY 0 1 VISIT 0 0 1 0
                         THERAPY*VISIT 0 0 0 0 0 0 1 0 BASVAL 17.857 BASVAL*VISIT 0 0 17.857 0;
  estimate 'Week 6 PLCB' INTERCEPT 1 THERAPY 0 1 VISIT 0 0 0 1
                         THERAPY*VISIT 0 0 0 0 0 0 0 1 BASVAL 17.857 BASVAL*VISIT 0 0 0 17.857;
  estimate 'Week 1 DIFF' THERAPY 1 -1 THERAPY*VISIT 1 0 0 0 -1 0 0 0;
  estimate 'Week 2 DIFF' THERAPY 1 -1 THERAPY*VISIT 0 1 0 0 0 -1 0 0;
  estimate 'Week 4 DIFF' THERAPY 1 -1 THERAPY*VISIT 0 0 1 0 0 0 -1 0;
  estimate 'Week 6 DIFF' THERAPY 1 -1 THERAPY*VISIT 0 0 0 1 0 0 0 -1;
run;
```

BGLIMM プロシジャを実行すると，収束診断の結果やプロット，MODEL 文で指定したパラメータおよび ESTIMATE 文で指定したパラメータの線形結合に対する事後分布の要約統計量や 95%最高事後密度（HPD）信用区間などが出力される。頻度流 MMRM と Bayesian MMRM の各時点の各群のベースラインからの変化量の調整済み平均値および群間差を図 1 に示す。無情報事前分布を用いていることから，頻度流 MMRM と Bayesian MMRM でほぼ同様の解析結果が得られた。なお，ベイズ流の解析では，パラメータがある区間に含まれる確率を求めることができ，Week 6 で PLACEBO 群と比べて DRUG 群のベースラインからの変化量が大きい確率（Week 6 の群間差の MCMC サンプルが 0 未満となる割合）は 99.4%であった。



（左）MIXED：調整済み平均±標準誤差，BGLIMM：調整済み事後平均±事後標準偏差

（右）MIXED：調整済み平均の群間差（95%信頼区間），BGLIMM：調整済み事後平均の群間差（95% HPD 区間）

図 1 MIXED プロシジャの頻度流 MMRM と BGLIMM プロシジャの Bayesian MMRM の結果

5. MCMC プロシジャによる Bayesian MMRM

MCMC プロシジャは，データの尤度関数やパラメータの事前分布，必要に応じて変量効果を指定することで，対応する事後分布から MCMC 法によるサンプルリングを行うベイズ流解析のための汎用的なプロシジャである。MCMC プロシジャは一連の標準的な分布をサポートしているだけでなく，データステップと同様のプログラミング文で任意の尤度関数や事前分布などを指定可能であり，広範なベイズ流の解析に柔軟に対応できる。一方で，BGLIMM プロシジャなどの特定のモデルに特化したプロシジャと異なり，使いこなすためにはベイズ流の解析の知識やプログラミング技術が必要である。

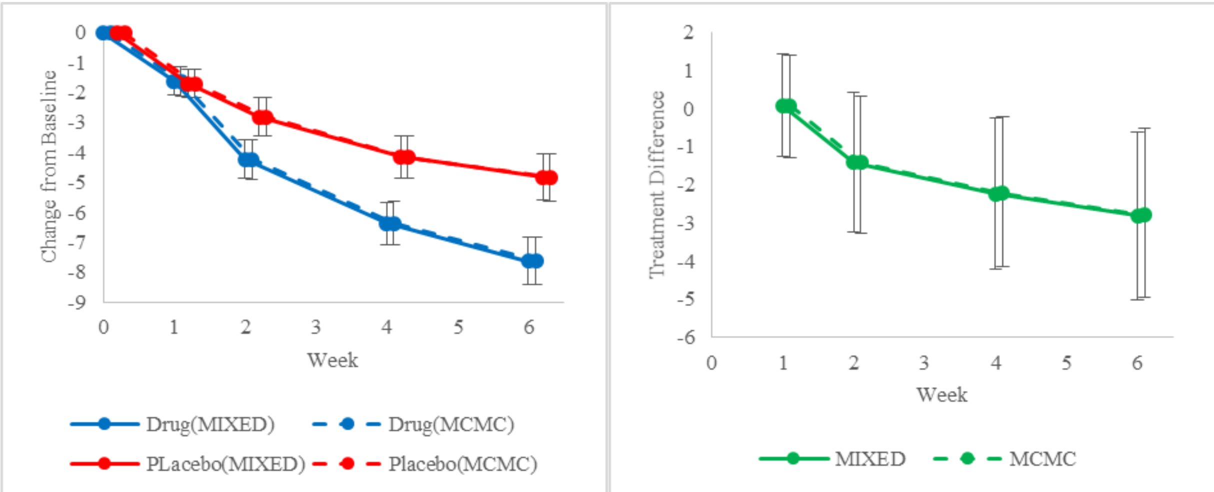
MCMC プロシジャを用いて，Bayesian MMRM による解析を行う SAS のコードを以下に示す。MCMC プロシジャで，被験者内の経時測定データ間の共分散構造をモデル化するためには，解析前に TRANSPOSE プロシジャなどを用いて同一被験者のすべての経時測定データが 1 オブザベーションに含まれるデータ構造に変換する必要がある。MCMC プロシジャでは，ARRAY 文を使って経時測定データに対応する変数を配列化し，MODEL 文で配列の変数が多変量正規分布（MVN）に従うことを指定する。多変量正規分布の平均および共分散行列も ARRAY 文を使って 1 次元および 2 次元の配列で定義してから指定する（固定効果パラメータなど他の変数も配列化しておくとプログラム上便利である）。モデルは 4 節と同じであり，平均構造に必要な固定効果パラメータは，切片（beta0），群（beta_t），ベースライン値（beta_b）で各 1 個，時点（beta_v4-beta_v6），群と時点の交互作用（beta_tv4-beta_tv6），ベースライン値と時点の交互作用（beta_bv4-beta_bv6）で各 3 個（時点数−1）の合計 12 個となる。パラメータとみなす変数は PARAMS 文で定義する。共分散パラメータは配列のまま指定することができる。PARAMS 文は複数指定可能であり，各 PARAMS 文で指定した変数がブロック化され，ブロックごとにサンプリングが行われる。PARAMS 文で定義したパラメータの事前分布を PRIOR 文で指定する。事前分布も 4 節と同じ設定とした。逆ウィシャート分布のスケールパラメータ（行列）として単位行列を指定するが，MCMC プロシジャの実行において設定の段階で一度だけ実行すれば良いので，BEGINCNST/ENDCNST 文で囲んでおく。平均構造のモデル指定は，データセットの変数と定義したパラメータを用いて，データステップと同様なプログラミング文で記載する。用いるパラメータは異なるが，記載は時点によらず共通のため，DO ループを使用するのが簡便である。MCMC プロシジャでは，調整済み事後平均（DRUG 群：lsm14-lsm17，PLACEBO 群：lsm24-lsm27）やその群間差（lsmd4-lsmd7）はプログラミング文で算出する。この計算はオブザベーションに依存しないため，MCMC の各反復の最初と最後にだけ実行される BEGINNODATA/ENDNODATA 文で囲む。各パラメータの係数は，BGLIMM プロシジャの ESTIMATE 文と同様である。PROC MCMC 文は，BGLIMM プロシジャと同様であり，バーンインやサンプリングの回数，出力データセットなどを指定する。MONITOR で指定した変数（_PARMS_は，PARMS 文で指定したパラメータを意味する）が，データセットへの出力および解析の対象になる。

```
proc transpose data=chapter15_example out=T prefix=CHG;
  var CHANGE;
  id VISIT;
  by PATIENT THERAPY BASVAL;
run;
```

```
proc mcmc data=T seed=1234 nbi=100000 nmc=500000 thin=100 outpost=POST monitor=(_parms_ lsm14-lsm17 lsm24-
lsm27 lsmd4-lsmd7) diag=all plot(smooth)=all;
  array y [4] CHG4-CHG7;
  array mu[4] mu1-mu4;
  array R[4, 4];
  array S[4, 4];
  array beta_v [4] beta_v4-beta_v6 0;
  array beta_tv[4] beta_tv14-beta_tv16 0;
  array beta_bv[4] beta_bv4-beta_bv6 0;
  array lsm1[4] lsm14-lsm17;
  array lsm2[4] lsm24-lsm27;
  array lsmd[4] lsmd4-lsmd7;

  begincnst;
    call identity(S);
  endcnst;

  parms beta0 beta_b 0;
  parms beta_v4 beta_v5 beta_v6 0 ;
  parms beta_t 0;
  parms beta_tv14 beta_tv15 beta_tv16 0;
  parms beta_bv4 beta_bv5 beta_bv6 0;
  parms R;

  prior beta: ~ normal(0, var=10000);
  prior R ~ iwish(4, S);

  beginnodata;
    do j=1 to 4;
      lsm1[j] = beta0 + beta_b*17.857 + beta_t + beta_v[j] + beta_tv[j] + beta_bv[j]*17.857;
      lsm2[j] = beta0 + beta_b*17.857          + beta_v[j]              + beta_bv[j]*17.857;
      lsmd[j] = beta_t + beta_tv[j];
    end;
  endnodata;

  do j=1 to 4;
    if THERAPY='DRUG'
      then mu[j] = beta0 + beta_b*BASVAL + beta_t + beta_v[j] + beta_tv[j] + beta_bv[j]*BASVAL;
      else mu[j] = beta0 + beta_b*BASVAL +         + beta_v[j]              + beta_bv[j]*BASVAL;
```

```
  end;

  model y ~ mvn(mu, R);
run;
```

MCMC プロシジャでも BGLIMM プロシジャと同様の出力が行われる。なお，MCMC プロシジャは，SAS/STAT 12.1 以降，デフォルトでは応答変数の欠測値に MAR を仮定し，欠測値をパラメータとみなして補完する。MCMC プロシジャによる Bayesian MMRM でも頻度流 MMRM とほぼ同じ結果が得られた（図 2）。また，Week 6 の群間差が 0 未満となる確率（Week 6 の群間差の MCMC サンプルが 0 未満となる割合）は 99.3%であった。



（左）MIXED：調整済み平均±標準誤差，MCMC：調整済み事後平均±事後標準偏差

（右）MIXED：調整済み平均の群間差（95%信頼区間），MCMC：調整済み事後平均の群間差（95% HPD 区間）

図 2 MIXED プロシジャの頻度流 MMRM と MCMC プロシジャの Bayesian MMRM の結果

6. 回帰モデルに基づく欠測値の補完

MMRM と同様に欠測メカニズムが MAR のもとで妥当な解析方法として，多重補完法がある（野間 2017）。多重補完法は，何らかの方法で欠測値を補完した複数個のデータセットを生成し，複数個のデータセットをそれぞれ解析して得られたパラメータ推定値とその標準誤差を統合し，最終的な解析結果を得る方法である。SAS では，MI プロシジャ（欠測値を補完したデータセットの生成），推定値を得るための解析プロシジャおよび MIANALYZE プロシジャ（解析結果の統合）で実行することができる。MI プロシジャで実装されている補完方法の中に，単調回帰（日本製薬工業協会 2016）および fully conditional specification 法（野間 2017）がある。これらの方法は，補完対象の変数を応答変数，補完に用いる変数を説明変数とした回帰モデルで欠測値を補完する。経時測定データの場合，例えば，補完対象の時点の測定値を応答変数，それ以前の時点の測定値（および有効性に影響する背景因子）を説明変数とする。

本節では，MCMCプロシジャで同様な回帰モデルに基づいて欠測値を補完する方法を紹介する。y_{ij}を被験者iのj時点目（$j = 1, \cdots, 5$）の測定値（これまでの節と異なり，本節ではベースラインからの変化量ではないことに注意）とする。ベースラインには欠測値がないものとし，それ以降の各時点（$j = 2, \cdots, 5$）の欠測値の補完に，以下の回帰モデルを用いる。

$$y_{ij} \sim N\left(\mu_{ij}, \sigma_j^2\right), \qquad \mu_{ij} = \begin{cases} \beta_{Pj0} + \beta_{Pj1} y_{i1} + \cdots + \beta_{Pj(j-1)} y_{i(j-1)} & : trt_i = 0 \\ \beta_{Tj0} + \beta_{Tj1} y_{i1} + \cdots + \beta_{Tj(j-1)} y_{i(j-1)} & : trt_i = 1 \end{cases}$$

ここで，$trt_i = \{0: \mathrm{PLACEBO}, 1: \mathrm{DRUG}\}$は群の指示変数であり，$\beta$の1つ目の添え字（PまたはT）は群，2つ目の添え字は応答変数（補完対象）の時点，3つ目の添え字は説明変数の時点を表している。また，σ_j^2は時点jの応答変数の分散である。そして，主な関心がある治療終了時点（Week 6）のベースラインからの変化量$d_i = y_{i5} - y_{i1}$に対して，以下の共分散分析モデルで解析を行う（他の時点についても解析可能である）。

$$d_i \sim N\left(\mu_i, \tau^2\right), \qquad \mu_i = \alpha + \delta trt_i + \varphi y_{i1}$$

ここで，αは切片，δは群，φはベースライン値に関するパラメータ，τ^2はd_iの分散である。各β，α，δおよびφの事前分布として，無情報の正規分布$N(0, 10^4)$，各σ^2およびτ^2の事前分布として，形状パラメータおよびスケールパラメータを0.01とした無情報の逆ガンマ分布$IG(0.01, 0.01)$を用いた。これらのモデルに含まれるパラメータおよびy_{ij}の欠測値（パラメータとみなされる）は，単一のMCMCによってサンプリングされる。なお，ベースライン値などの背景因子に欠測値がある場合でも，それらの変数を補完するための回帰モデルを追加することで，背景因子の欠測値を補完して解析することが可能である。

SASのコードを以下に示す。5節と同様，事前にデータセットをTRANSPOSEプロシジャで転置する。MCMCプロシジャでは，複数のMODEL文を指定することができ，各時点の補完モデルおよび治療終了時点のベースラインからの変化量に対する共分散分析モデルを，各MODEL文で指定している。5節でも述べたように，応答変数の欠測はMARの仮定で補完される。共分散分析のMODEL文で，他のMODEL文と同様にCHG ~ NORMAL(MU, VAR=TAU)の指定を行うと，「The procedure is modeling missing values in the response variable CHG, therefore it cannot have values assigned to it in the program.」というエラーが生じて実行できないため，プログラムで定義した一般の対数尤度を指定することができ，応答変数を指定する必要がないGENERAL関数を用いている。LPDFNORM(X, MU, SD)は，MCMCプロシジャ内で使用できる関数で，正規分布の確率密度関数の対数変換値を返す。5節と同様に，BEGINNODATAとENDNODATAで囲んだ部分で，各群のベースラインからの変化量の調整済み平均を算出している（群間差はδである）。

```
proc transpose data=chapter15_example out=T(rename=(BASVAL=Y1 _4=Y2 _5=Y3 _6=Y4 _7=Y5));
  var HAMDTL17;
  id VISIT;
  by PATIENT THERAPY BASVAL;
run;

proc mcmc data=T seed=123456 nbi=100000 nmc=500000 thin=100 outpost=POST monitor=(_parms_ lsm1 lsm2)
diag=all plot(smooth)=all;
  parms beta_p20 beta_p21 beta_t20 beta_t21 sigma2;
  parms beta_p30 beta_p31 beta_p32 beta_t30 beta_t31 beta_t32 sigma3;
```

```
  parms beta_p40 beta_p41 beta_p42 beta_p43 beta_t40 beta_t41 beta_t42 beta_t43 sigma4;
  parms beta_p50 beta_p51 beta_p52 beta_p53 beta_p54 beta_t50 beta_t51 beta_t52 beta_t53 beta_t54 sigma5;
  parms alpha delta phi tau;

  prior beta: alpha delta phi ~ normal(0, var=10000);
  prior sigma: tau ~ igamma(shape=0.01, scale=0.01);

  if THERAPY='PLACEBO' then mu2 = beta_p20 + beta_p21 * Y1;
                       else mu2 = beta_t20 + beta_t21 * Y1;
  if THERAPY='PLACEBO' then mu3 = beta_p30 + beta_p31 * Y1 + beta_p32 * Y2;
                       else mu3 = beta_t30 + beta_t31 * Y1 + beta_t32 * Y2;
  if THERAPY='PLACEBO' then mu4 = beta_p40 + beta_p41 * Y1 + beta_p42 * Y2 + beta_p43 * Y3;
                       else mu4 = beta_t40 + beta_t41 * Y1 + beta_t42 * Y2 + beta_t43 * Y3;
  if THERAPY='PLACEBO' then mu5 = beta_p50 + beta_p51 * Y1 + beta_p52 * Y2 + beta_p53 * Y3 + beta_p54 * Y4;
                       else mu5 = beta_t50 + beta_t51 * Y1 + beta_t52 * Y2 + beta_t53 * Y3 + beta_t54 * Y4;
  if THERAPY='PLACEBO' then mu = alpha + phi * Y1;
                       else mu = alpha + delta + phi * Y1;

  beginnodata;
    lsm1 = alpha + phi * 17.857;
    lsm2 = alpha + delta + phi * 17.857;
  endnodata;

  model Y2 ~ normal(mu2, var=sigma2);
  model Y3 ~ normal(mu3, var=sigma3);
  model Y4 ~ normal(mu4, var=sigma4);
  model Y5 ~ normal(mu5, var=sigma5);

  CHG = Y5 - Y1;
  ll = lpdfnorm(CHG, mu, sqrt(tau));
  model general(ll);
run;
```

表 3 欠測値を補完した共分散分析の解析結果

群	DRUG		PLACEBO		群間差	95%HPD 区間	
時点	調整済み事後平均	事後標準偏差	調整済み事後平均	事後標準偏差		下限	上限
Week 6	-7.79	0.73	-5.07	0.70	-2.72	-4.69	-0.75

共分散分析モデルによる解析で得られた各群のベースラインからの変化量の調整済み平均およびその群間差を表 3 に示す。本節の解析でも，頻度流 MMRM と一貫した結果が得られた。

7. その他のトピックとまとめ

7.1 BGLIMM プロシジャと MCMC プロシジャの比較

BGLIMM プロシジャおよび MCMC プロシジャによる Bayesian MMRM でほぼ同じ結果が得られ，本稿では無情報事前分布を用いたことから，MIXED プロシジャによる頻度流 MMRM とも同様の結果であった。

一般化線形混合効果モデルは広範なモデルを含んでおり，BGLIMM プロシジャの適用範囲は広いが，MCMC プロシジャはより広範なモデルを扱うことができる汎用的なプロシジャである。MCMC プロシジャにしかできないこととして，例えば，正規分布に従わない変量効果，BGLIMM プロシジャでは指定できない事前分布，6 節のような複数の MODEL 文を用いた解析および欠測メカニズムに missing not at random を仮定した解析などが考えられる。

一方で，MCMC プロシジャは，特定のモデルに合わせてカスタマイズされていない一般的なメトロポリス・ヘイスティングス法を含むサンプリングアルゴリズムを使用するが，BGLIMM プロシジャはモデルに特化した最適なアルゴリズムによる効率的なサンプリングを使用する。本稿の Bayesian MMRM の実行でも，BGLIMM プロシジャではすべてのパラメータに対して共役事前分布を利用したギブスサンプリング法による効率的なサンプリングが行われており，バーンインおよびサンプリング回数は少なくても問題ないと考えられた。一方で，MCMC プロシジャでは一部のパラメータで高い自己相関が認められ，BGLIMM プロシジャよりもバーンインおよびサンプリング回数を増やす必要があると考えられた。そのため，どちらのプロシジャでも扱えるモデルの場合，BGLIMM プロシジャを使用する方が効率的と考えられる。

MCMC プロシジャでは，プログラミングでモデルパラメータの線形結合のみならず，パラメータ変換（例えば，指数変換）なども行うことができる。BGLIMM プロシジャでは，ESTIMATE 文でパラメータの線形結合を行うことはできるが，パラメータ変換は行うことができず，OUTPOST でデータセットに出力した後，データステップでパラメータ変換をする必要がある。ただし，OUTPOST で出力したデータセットに対して，事後要約統計量の算出，収束診断や各種プロットの作成を行うための AUTO CALL マクロを利用できる（MCMC プロシジャでも利用できる）。例えば，%SUMINT（事後要約統計量および 95% HPD 区間の算出）や%TADPLOT（トレースプロット，自己相関プロット，密度プロットの作成）などがある。詳細は，SAS/STAT User’s Guide の BGLIMM プロシジャおよび MCMC プロシジャの章（SAS Institute Inc. 2023）を参照されたい。

7.2 情報のある事前分布

本稿のベイズ流の解析では，すべて無情報事前分布を用いた。ベイズ流アプローチの特徴の 1 つは，治療効果などに関する既存情報を事前分布として表現し，解析で考慮できることである。情報のある事前分布を用いる場合でも，MCMC プロシジャでは PRIOR 文の事前分布の指定に事前情報を反映すれば良い。BGLIMM プロシジャでは，COEFFPRIOR=NORMAL(INPUT=データセット名)で，事前分布として用いる（多変量）正規分布のパラメータ（平均，分散または共分散行列）を読み込む必要がある。データセットの変数には，事前分布を指定したい MODEL 文の変数，_TYPE_，_NAME_（_TYPE_=’COV’の場合）を含める。事前分布を指定する変数を X1，X2 とした場合のデータセットの例を図 3 に示す。データセットに含ま

れていない変数の事前分布には，無情報の正規分布が用いられる。CLASS 文で指定した変数や交互作用項に含まれるパラメータの変数名は，OUTPOST の出力データセットの変数名を参照すれば良い。

TYPE	X1	X2
MEAN	99	99
VAR	99	99

TYPE	_NAME_	X1	X2
MEAN		99	99
COV	X1	99	99
COV	X2	99	99

図 3 情報のある事前分布を指定するデータセットの例（左：平均と分散，右：平均と共分散行列）

本稿では，Bayesian MMRM の固定効果に，頻度流 MMRM の固定効果に通常含まれる群，時点，群と時点の交互作用などを含めた。そのため，各時点の各群の平均および群間差は，複数の固定効果パラメータの線形結合で表される。このことは，無情報事前分布の場合は特に問題ないと考えられるが，各群の平均や群間差に既存情報を考慮したい場合，各群の平均や群間差自体をパラメータとしてモデルに含め，その事前分布に情報のある事前分布を設定できる方が良いと考えられる。本稿で用いたモデルと本質的には変わらないが，パラメータ構成が異なるモデルを考えることができる（$\mathbf{X}_i$と$\boldsymbol{\beta}$は異なるが，$\mathbf{X}_i\boldsymbol{\beta}$は同じ）。例えば，本稿のモデルから切片，群，時点，ベースライン値を除き，群と時点の交互作用，ベースライン値と時点の交互作用のみ含めたモデルでは，群と時点の交互作用は，（共変量がある場合はその値が 0 のときの）各時点の各群の平均に対応する。変更した場合の MODEL 文と ESTIMATE 文（Week 6 のみ示す）は，以下のようになる。

```
model CHANGE = THERAPY*VISIT BASVAL*VISIT /noint dist=normal coeffprior=normal;
estimate 'Week 6 DRUG' THERAPY*VISIT 0 0 0 1 0 0 0 0 BASVAL*VISIT 0 0 0 17.857;
estimate 'Week 6 PLCB' THERAPY*VISIT 0 0 0 0 0 0 0 1 BASVAL*VISIT 0 0 0 17.857;
estimate 'Week 6 DIFF' THERAPY*VISIT 0 0 0 1 0 0 0 -1;
```

また，時点，群と時点の交互作用，ベースライン値と時点の交互作用のみ含めたモデルでは，群と時点の交互作用は，（共変量の有無を問わず）各時点の群間差に対応する。変更した場合の MODEL 文と ESTIMATE 文（Week 6 のみ示す）は，以下のようになる（群間差は群と時点の交互作用を参照すれば良く，ESTIMATE 文で指定する必要はない）。

```
model CHANGE = VISIT THERAPY*VISIT BASVAL*VISIT /noint dist=normal coeffprior=normal;
estimate 'Week 6 DRUG' VISIT 0 0 0 1 THERAPY*VISIT 0 0 0 1 0 0 0 0 BASVAL*VISIT 0 0 0 17.857;
estimate 'Week 6 PLCB' VISIT 0 0 0 1 THERAPY*VISIT 0 0 0 0 0 0 0 1 BASVAL*VISIT 0 0 0 17.857;
```

Bayesian MMRM では，すべての時点の平均や群間差がモデル化されているが，主な関心がある治療終了時点以外の時点に関する事前情報を入手したり，解析で考慮したりすることが困難な場合があると想定される。6 節で紹介した回帰モデルによって欠測値を補完する方法では，主な関心がある治療終了時点以外の時点のデータは欠測値の補完のみに利用され，最終的な解析モデルは治療終了時点の平均や群間差のみモデル

化している。治療終了時点のみ事前情報を考慮したい状況では，欠測値を補完する方法の方が利用しやすいと考えられる。

7.3 まとめ

本稿では，BGLIMM プロシジャおよび MCMC プロシジャを用いて，ベイズ流の経時測定データ解析方法である Bayesian MMRM および回帰モデルによって欠測値の補完を行う方法を紹介した。なお，本稿では解析の目的が治療終了時点の平均値の群間比較である状況を想定したため取り上げなかったが，経時測定データに対する他の解析方法として，時点を連続変数とし，被験者ごとの切片およびスロープのばらつきを変量効果とした線形混合効果モデルなどもあり（Dmitrienko and Koch 2017），そのような解析も本稿で紹介したプロシジャで対応可能である。

本稿が SAS でベイズ流の経時測定データ解析を実施する際の参考になると幸いである。

参考文献

Biglan K, Munsie L, Svensson KA, Ardayfio P, Pugh M, Sims J, Brys M (2022). Safety and Efficacy of Mevidalen in Lewy Body Dementia: A Phase 2, Randomized, Placebo-Controlled Trial. *Mov Disord*. 37(3):513-524.

Drug Information Association Scientific Working Group on Estimands and Missing Data. https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data#dia-missing-data [2023/8/29 アクセス]

Dmitrienko A, Koch G (2017). Analysis of Clinical Trials Using SAS (Second Edition). SAS Press: Cary, NC, USA.

Chen F, Brown G, Strokes M (2016). Fitting your favorite mixed models with PROC MCMC. https://support.sas.com/resources/papers/proceedings16/SAS5601-2016.pdf [2023/8/29 アクセス]

Food and Drug Administration (2019). Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products. https://www.fda.gov/media/130897/download [2023/8/29 アクセス]

Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA (2004). Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. *J Clin Psychopharmacol*. 24(4):389-399.

Hirakawa, A, Sato, H, Igeta, M, Fujikawa, K, Daimon, T, Teramukai, S (2022). Regulatory issues and the potential use of Bayesian approaches for early drug approval systems in Japan. *Pharmaceutical Statistics* 21(3): 691-695.

Ivy D, Bonnet D, Berger RMF, Meyer GMB, Baygani S, Li B; LVHV Study Group (2021). Efficacy and safety of tadalafil in a pediatric population with pulmonary arterial hypertension: phase 3 randomized, double-blind placebo-controlled study. *Pulm Circ*. 11(3):1-8.

SAS Institute Inc. (2023). SAS/STAT 15.3 User’s Guide. SAS Institute Inc., Cary, NC, USA.

厚生省医薬安全局審査管理課長 (1998).「臨床試験のための統計的原則」について. https://www.pmda.go.jp/files/000156112.pdf [2023/8/29 アクセス]

五所 正彦, 丸尾 和司 (2017). 経時測定データ解析における mixed-effects models for repeated measures（MMRM）の利用. 応用統計学 46(2):53-65.

日本製薬工業協会 (2016). 欠測のある連続量経時データに対する統計手法について Ver2.0. https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc0000007qqq-att/statistics01.pdf [2023/8/29 アクセス]

野間 久史 (2017). 連鎖方程式による多重代入法. 応用統計学 46(2):67-86

ベイズパターン認識によるグラフ品質管理の自動化

○福島　綾介

（イーピーエス株式会社）

Automated Graph Quality Control with Bayesian Pattern Recognition

Fukushima, Ryosuke

EPS Corporation

要旨

臨床試験の統計解析結果はグラフで可視化されることが多く、その品質管理のためにダブルプログラミングが採用される。そのダブルプログラミングで得た 2 つの結果を照合することで、ミスを発見し品質の向上が実現できる。しかし、解析結果のグラフは最終成果物が画像であるため、目視での照合作業となる場合が多く、時間がかかり、照合作業の質は担当する人によって異なる。

本発表では、グラフの照合作業をベイズパターン認識で自動化する試みを紹介する。ベイズパターン認識は SAS の Proc MCMC により実装可能であり、グラフ画像上の座標軸を識別し、プロット点の位置を読み出すことができる。目視によるグラフのプロット点読み出しよりも、短時間で高精度に処理を行えるため、品質管理業務の効率化が見込める。

キーワード：グラフ品質管理、グラフ照合作業、パターン認識、ベイズ統計、業務自動化

序論

品質管理のためダブルプログラミングと照合作業が行われる

臨床試験において解析結果の品質管理のためにダブルプログラミングが行われる。ダブルプログラミングでは 2 人の作業者で同様の作業を独立に行い結果を照合する。結果の照合により、プログラムのミスを発見し修正することで、品質の向上が可能になる。この照合作業において、解析結果が表であるときは、表内の対応する行と列にある数値や文字列が一致するか簡単に比較できる。しかし、解析結果がグラフであるときは、その比較は表ほど簡単にはいかない。これは解析結果が画像フォーマットで出力されるためである。

グラフの照合作業は目視となり手間がかかる

以下では、解析結果が散布図のピクセル画像フォーマットとして出力されるときの照合作業について考える。この照合作業で比較対象となるのは、グラフのプロット位置をグラフ上の座標位置から読みだした数値である。しかし、この比較は画像同士の単純な比較では実現できない。これはピクセル画像フォーマットには色の濃淡の情報は含まれていても、画像上のどの直線が座標軸であるのかといった情報が直接的には含まれていないからである。そのため、画像を目視で確認しプロット位置を読み出す必要がある。

実際の照合作業においては、片方の担当者がグラフを画像として出力するのに対して、もう片方の担当者は散布図プロットの座標値をテキストファイルとして出力する。この状況下の比較であれば、画像を目視で確認しプロット位置を読み出し、テキストファイルの数値と一致するかどうか比較できる。しかし、散布図においてプロット点が数 100 や数 1000 を超える場合は照合するプロット点を絞ることが多い。これは目視での照合に時間がかかりすぎてしまうからである。

目視照合作業の課題として以下の 3 つが挙げられる。(1) 作業に時間がかかる。(2) 目視作業によるプロット点の区別には限界がある。(3) 人によって作業の精度が異なる。(4) 照合作業結果の記録が難しい。(1) は目視作業であることに加えて、グラフ仕様の修正があった際には照合作業を繰り返すことが求められ、同様の体裁のグラフを異なる変数や条件などで作成した場合には、さらに作業量が増える。(2) はプロット点が少なく、まばらに散っている場合は問題にならない。しかし、複数のプロット点が近くに寄った場合、目視では異なるプロット点を区別できなくなる。(3) は人が手作業を行う以上避けることはできず、疲労度合いによっても作業の精度は変わりうる。(4) は目視で照合作業を行い、一致が確認できたとしてもその証跡を残すことは難しい。仮に照合作業を行わなくても一致したという記録を残せる。

本論文で提供する方法はグラフ照合作業の自動化に活用できる

本論文では、グラフ画像内に保存された情報にアクセスし、グラフ画像のプロット位置をグラフ上の座標位置から自動で読み出すプログラムの作成方法を提供する。この方法はグラフの目視照合作業において、以下のような目標を達成できる。(1') 作業時間の短縮。(2') 目視では判別できない重なり合ったプロット点の区別。(3') 作業の高精度化 (4')　照合作業結果の電子的記録。(1') は照合作業にかかる時間が短縮されるため、繰り返し作業が必要でも効率化できる。これによって、プロット点が多く、目視作業で全てのプロット点を照合することを妥協せざるを得ない状況でも、全てのプロット点を照合できる。(2') は個々のプロット点の座標を自動で読み出すことで、画像上で重なり合ったプロット点を個別に読み出せる。(3') は画像内に

保存された情報にアクセスするため目視での読み出しよりも高精度化でき、プログラムによる機械的な作業になるため状況による精度の変化が少なく済む。(4’) は自動化プログラムにより、実行日、解析対象の画像、比較結果を適切にファイルとして出力することで、照合作業結果の電子的記録を残すことができる。

本論文で提供する方法は、グラフ画像からプロット位置を読み出すために、以下の手順で画像情報の処理を行う。(A) グラフ画像から画像情報の抜き出し (B) グラフ構成要素のパターン認識 (C) プロット位置の読み出し。本方法では、グラフ照合作業におけるグラフ画像フォーマットは Enhanced Metafile Format (EMF)画像を対象とする。EMF 画像はベクター画像フォーマットであり、画像上の特定の場所に点や直線を配置する命令をレコードとして直接保持するデータ形式である。そのため、適切な情報処理によって、画像に含まれる全ての直線がそれぞれ画像上のどの点からどの点までに引かれているのか正確に知ることができる。(A) では C++によるプログラムを記述しレコード情報を読み出す。しかし、画像に含まれる直線の内、どれが X 軸なのか、どれが軸目盛なのか知ることはできない。そのため、(B) でベイズパターン認識を用いた EMF 画像のレコード情報の識別を行う。これによって、画像内の直線の内、どれが X 軸であるのか軸目盛であるのか予測できる。最後に(C)でパターン認識によって得られた目盛ラベルの情報からプロット位置を読み出す。

本論文では散布図から正確にプロット位置を読み出せることを示す

どんな種類のグラフをどんな記法で作成しても本論文で示す方法が活用できることが理想であるが、本論文では散布図を Proc SGPLOT の Scatter statement によって作成したものをサンプル画像として解析を行う。この解析結果は、散布図を作成した元データと比較して、十分な精度でプロット位置を読み出せており、その絶対誤差は最大でも 1E-3 程度であった。今回用意したサンプル画像に対して、絶対誤差 1E-3 程度の精度でプロット位置を目視で数値読み出しするのは不可能である。

本論文で提供する方法は臨床試験の結果をグラフとして示す場合に、ダブルプログラミングで得られた結果の照合作業を自動化するために活用できる。従来の目視照合よりも短時間で高精度に検証できるため、高度な品質管理が可能になる。現時点で適用可能なグラフの種類は散布図に限られるが、棒グラフや箱ひげ図、カプラン・マイヤー曲線などにも適用可能であり、臨床試験において高頻度に作成されるグラフについては作業の効率化が望める。

方法

サンプル画像の作成

サンプルとなる EMF 画像は SAS で 2 次元正規乱数を生成し、Proc SGPLOT を用いて散布図を作成し、640×480 ピクセルの EMF 画像として保存した。2 次元正規乱数は X 軸方向の平均と分散はそれぞれ 5 と 2、Y 軸方向の平均と分散はそれぞれ 50 と 20 とし、100 点分の乱数生成を行った。

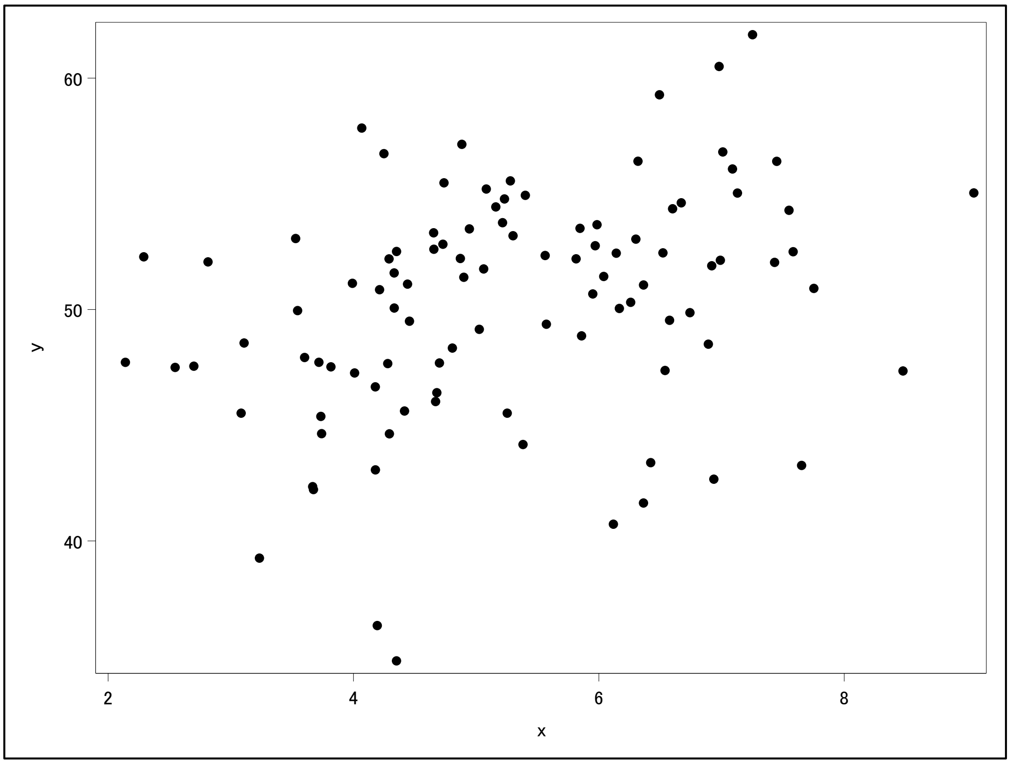


図 1 : 解析対象となるサンプル画像

EMF 画像からのレコード情報の読み出し

EMF 画像は SAS の ods graphics のオプションから imagefmt = EMF を指定することで容易に作成可能である。その EMF 画像のデータ仕様は Microsoft 社から公開されており [1]、データ仕様さえ理解できればレコード情報を読み出すことができる。本論文では、このレコード情報の読み出しに利用可能な Windows Application Programming Interface (Windows API)を利用するため、Microsoft Visual Studio を用いた C++プログラムを作成した。Windows API の中でも EnumEnhMetaFile 関数と PlayEnhMetaFileRecord 関数を組み合わせることで、EMF 画像のレコードを先頭から 1 つずつ順番に扱うことができるため、開発が容易になる。最終的に EMF 画像のレコードをテキストファイルとして出力した。

パターン認識前の前処理

EMF 画像を直接テキストファイルに変換したままでは、この後の処理であるベイズパターン認識には適していないため、データ整形を SAS で行った。この前処理は SAS の DATA ステップがプログラムデータベクトル（PDV）に基づいて行う処理と相性が良い。以下で、その理由と EMF 画像レコードの特徴について記述する。

散布図のプロット位置読み出しには、X 軸などの直線を描くレコード群と目盛ラベルなどの文字列を描くレコード群を抽出し、画像上で描画される座標位置を得る必要がある。しかし、EMF 画像のレコードは画像上で描画される座標位置をそのまま保持しているわけではない。例えば、1 本の直線を描く場合以下の 4 種のレコードに情報が分かれて記録されている。それらのレコードは、直線オブジェクトを描く座標を記録するオブジェクトレコード、線種や色等を記録するペンレコード、オブジェクトレコードとペンレコードを指定し直線を描く描画レコード、オブジェクトの座標をアフィン変換（平行移動と拡大縮小、回転などを行う変換）し実際の EMF 画像上の座標に変換する変換レコードである。このようなレコードが順番に実行され画像

上に描画されるが、複数の直線を描く場合などではオブジェクトレコードやペンレコードを上書き更新するレコードが作成される。そのため、画像上で描画される直線の座標位置を得るためには、直線の描画レコードが実行されるよりも前の最後に実行されたオブジェクトレコード、ペンレコード、変換レコードを関連付けて抽出する必要がある。この処理は SAS であれば DATA ステップで retain ステートメントを用いて、オブジェクトレコード、ペンレコード、変換レコードの状態を記録し、描画レコードが現れたときに output ステートメントで出力すれば良く、簡潔に処理できる。これは描画レコードが処理を行うときのオブジェクトレコードやペンレコードの状態に依存し、それより以前のレコードの状態に依存しないためである。

上記の方法により、描画されるオブジェクトの画像上での座標位置が正しく抽出されているか確認するために Microsoft Paint を用いた。Microsoft Paint で EMF 画像を開き、カーソルを特定の位置に合わせると、そのカーソルのある画像上での座標位置がウィンドウ左下に表記される。これと比較することで座標位置が正しく抽出されていることを確認した。

描画オブジェクトのベイズパターン認識

EMF 画像から読みだしたレコード情報の離散的カテゴリ分類をベイズパターン認識で行う。本論文では表 1 に記載のように描画オブジェクトを計 7 種のカテゴリへ分類した。散布図のグラフ上の座標値の読み出しに必要なものは、クラス 1 を除いた 6 種のカテゴリに属する描画オブジェクトの内、X 軸と Y 軸に加えて、軸目盛と目盛ラベルはそれぞれの軸で両端にあるもので十分である。

表 1 : 識別するレコードの分類

クラス	レコードの分類
クラス 1	識別不能 （下記のどれにも当てはまらない）
クラス 2	X 軸
クラス 3	Y 軸
クラス 4	X 軸目盛
クラス 5	Y 軸目盛
クラス 6	X 軸目盛ラベル
クラス 7	Y 軸目盛ラベル

以下でベイズパターン認識は何をしているのか概説する。本論文で実装するベイズパターン認識は最終的にある座標情報のデータ群が得られたときに、それぞれのデータがどのクラスに分類されるのか、その確率を返す。この確率を求めるためにベイズの定理を用いると、あるクラスに属する描画オブジェクトが持つ座標情報が取りうる値の確率と、データを得る前にあるクラスに属する確率を定める必要がある。前者は尤度関数であり、後者は事前分布である。本論文における尤度関数の役割は端的に言えば、確率的な座標情報の乱数生成であり、異なるグラフにおける X 軸の画像上での描画位置や長さの違いは乱数の生成によるものだと考える。事前分布の役割は座標情報の属するクラスの乱数生成である。ベイズパターン認識で行うことは、これらの尤度関数と事前分布を組み合わせることで、データの乱数生成を行い、実際に得られたデータとの

当てはまり具合が良いものに高い確率を割り当てて返すことである。事前分布に基づいてクラスの組み合わせを乱数生成し、それぞれのクラスのもつ座標情報を尤度関数で定めた確率分布に基づいて乱数生成する。このようにして得られた乱数の列と実際に得られたデータとの当てはまり具合から分類されるクラスの確率を評価する。良いモデルが構築できているならば、乱数生成から何度も繰り返すことで、いつかは当てはまりの良いクラスの組み合わせが探索できるはずである。これがベイズパターン認識で行われることである。

以下で本論文において実装したベイズパターン認識について記述する。EMF 画像が全部でN個のレコードをもつとき、n番目のレコードに含まれる描画オブジェクトを配置する座標情報をベクトルで表し$\boldsymbol{D_n}$と表記する。例えば、任意の正の実数をXとYと表記すると、点座標は$\boldsymbol{D_n} = (X, Y)$と表現できる。また、直線を描く描画オブジェクトは 2 点を指定することで$\boldsymbol{D_n} = (X_1, Y_1, X_2, Y_2)$と表現できる。ここで、1 点目の座標を$X_1$と$Y_1$、2 点目の座標を$X_2$と$Y_2$で表記した。次に$n$番目のレコードが$M$種類のカテゴリのうちどれに当てはまるかを整数の確率変数C_nで表す。例えば、n番目のレコードがクラス 2 である X 軸であるとき、$C_n = 2$とする。

N個のレコード全体の座標情報を$\boldsymbol{D}$、レコード全体の当てはまるカテゴリを$\boldsymbol{C} = (C_1, C_2, \dots, C_N)$とすれば、ベイズパターン認識で求める事後分布を$P(\boldsymbol{C}|\boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{\pi})$と表すことができ、以下のようにして求めることができる。

$$P(\boldsymbol{C}|\boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{C})P(\boldsymbol{C}|\boldsymbol{\pi})}{P(\boldsymbol{D})}$$

ここで、$\boldsymbol{\theta}$は尤度関数に必要なパラメータ全体であり、$\boldsymbol{\pi}$は事前分布に必要なパラメータ全体である。

尤度関数$P(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{C})$は以下のように定める。

$$P(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{C}) = \prod_{m=1}^{M}\prod_{n=1}^{N} P_m(\boldsymbol{D}_n|\boldsymbol{\theta}_m)^{z_{m,n}}$$

ここで、$P_m(\boldsymbol{D}_n|\boldsymbol{\theta}_m)$はクラス$m$である描画オブジェクトが$\boldsymbol{D_n}$をもつ確率を表す確率分布であり、パラメータ$\boldsymbol{\theta}_m$をもつ。$z_{m,n}$はダミー変数であり、$z_{m,n} = \{0,1\}$かつ$\prod_{m=1}^{M} z_{m,n} = 1$を満たすような One-hot エンコーディングを実現する変数である。パラメータ$\boldsymbol{\theta}_m$に対して事前分布$P(\boldsymbol{\theta}_m)$を与えることもできるが、本論文では$\boldsymbol{\theta}_m$を定数として事前に与えることにする。

以下で特定のクラスが座標情報$\boldsymbol{D_n}$をもつ確率分布を定めるが、紙面の都合上、クラス 2 のみ記載する。クラス 2 である X 軸が EMF 画像上（640×480）の 2 点の座標$\boldsymbol{D_n} = (X_1, Y_1, X_2, Y_2)$をもつ確率を以下のように表す。

$$P_2(\boldsymbol{D}_n|\boldsymbol{\theta}_2) = \mathrm{FBe}(X_1|\alpha_1, \beta_1, l_1, u_1)\mathrm{FBe}(Y_1|\alpha_2, \beta_2, l_2, u_2)\mathrm{FBe}(X_2 - X_1|\alpha_3, \beta_3, l_3, u_3)\mathrm{Del}(Y_2 - Y_1)$$

ここで、$\mathrm{FBe}(X|\alpha, \beta, l, u)$と$\mathrm{Del}(X)$は Four-parameter beta distribution とデルタ分布である[補足資料 1]。Four-parameter beta distribution はベータ分布の拡張された確率分布であり、パラメータlを最小、uを最大とする実数値をとる。αとβは確率分布の形状を定め。適切にパラメータを定めることで、直線の片方の点が 640×480

の EMF 画像上にあり、X 軸方向の長さが$X_2 - X_1$であり、Y 軸方向の長さが 0 であるような X 軸が現れる確率を定められる。本論文におけるパラメータの値は[補足資料 2]に記載した。

事前分布$P(\boldsymbol{C}|\boldsymbol{\pi})$は以下のように定める。

$$P(\boldsymbol{C}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \mathrm{Cat}(C_n|\boldsymbol{\pi})$$

ここで、$\mathrm{Cat}(C_n|\boldsymbol{\pi})$はカテゴリカル分布である[補足資料 1]。上記に加えて、SAS での実装上は X 軸と Y 軸はそれぞれ 1 つになるように制限を加えた。

以上のようにして定めた事後分布$P(\boldsymbol{C}|\boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{\pi})$を Proc MCMC で実装できるマルコフ連鎖モンテカルロ法（MCMC）を用いて求める。上記のような単純な定式化であれば、必ずしも MCMC を用いる必要はないが、パラメータ$\boldsymbol{\theta}_m$に対して連続的な事前分布を定めるような場合は MCMC の利用が必要となる。SAS の Proc MCMC 以外に、他にベイズパターン認識を実装できるソフトウェアには実行環境に R を要求する RStan [2] があるが、基本的に推定する確率変数は連続変数である必要がある。そのため、RStan は本論文における離散的カテゴリ分類には用いることができない。一方で、Proc MCMC は確率モデルが連続変数と離散変数の両方を含んだとしても実装可能であるため、本論文では SAS の Proc MCMC を採用した。

結果

パターン認識により適切にレコードを識別できた

表 2 はサンプルとして作成した EMF 画像から、直線を描くオブジェクトの座標情報を読み出したものであり、ベイズパターン認識によって識別されたクラスを記載している。EMF 画像上の座標情報は左上を$(0, 0)$とする座標となっており、EMF 画像の右下が$(640, 480)$となる。レコード 1 は X 軸を描くレコードであるが、その識別結果は X 軸であるクラス 2 とはなっていない。しかし、レコード 5 がレコード 1 と同一の座標情報を持っており、レコード 5 は X 軸と識別されているため、問題とはならない。SAS の SGPLOT を用いて EMF 画像を作成すると、X 軸と Y 軸は 2 度重ねて描かれるようであり、今回の方法では X 軸と Y 軸が 1 つだけ識別されるように制限したため、このような結果となったと思われる。他のレコードについても同様に問題なく、クラスの識別が行われていることが確認できた。

グラフの読み出し数値は目視より高精度であった

表 3 は散布図の元データである 2 次元正規乱数の値と、EMF 画像から読み出されたグラフ座標上のプロット位置の間で、X 軸方向と Y 軸方向の絶対誤差と直線距離を求めたものである。絶対誤差は元データの座標と読み出し座標の間の X 座標または Y 座標のみに注目したときの差であり、直線距離は 2 点間の最短距離である（図 2)。絶対誤差と直線距離は値が小さいほど正確な読み出しができることを意味する。X 軸と Y 軸方向の絶対誤差は平均でそれぞれ 2.94E-3 と 1.89E-6 であった。目視で数値を読み出した場合、精度は目盛間隔に依存することになる。ここで、目盛間隔の 10 分の 1 まで読み出すとすれば、X 軸と Y 軸方向の絶対誤差はそれぞれ最大で 0.2 と 1 となる。目視での絶対誤差を基準に比を計算すると、X 軸と Y 軸方向の絶対誤差

比はそれぞれ、1.47E-2 と 1.89E-6 である。EMF 画像のレコード情報からグラフのプロット位置を読み出すことで、絶対誤差比の小さい X 軸方向であっても、目視より 100 倍程度精度が高く読み出すことができた。

表 2：パターン識別結果

レコード	座標情報				識別されたクラス	
n	X_1	Y_1	X_2	Y_2	C	分類
1	58	427	629	427	1	識別不能
2	629	427	629	10	1	識別不能
3	58	10	629	10	1	識別不能
4	58	427	58	10	3	Y 軸
5	58	427	629	427	2	X 軸
6	66	427	66	432	4	X 軸目盛
7	223	427	223	432	4	X 軸目盛
8	381	427	381	432	4	X 軸目盛
9	538	427	538	432	4	X 軸目盛
10	58	427	58	10	1	識別不能
11	58	342	53	342	5	Y 軸目盛
12	58	194	53	194	5	Y 軸目盛
13	58	46	53	46	5	Y 軸目盛

表 3：読み出し結果の精度

	平均	標準偏差	最小	最大
絶対誤差（X）	2.94E-3	1.14E-6	2.93E-3	2.94E-3
絶対誤差（Y）	1.89E-6	2.41E-6	-2.41E-6	1.07E-5
直線距離	2.94E-3	1.14E-6	2.93E-3	2.94E-3

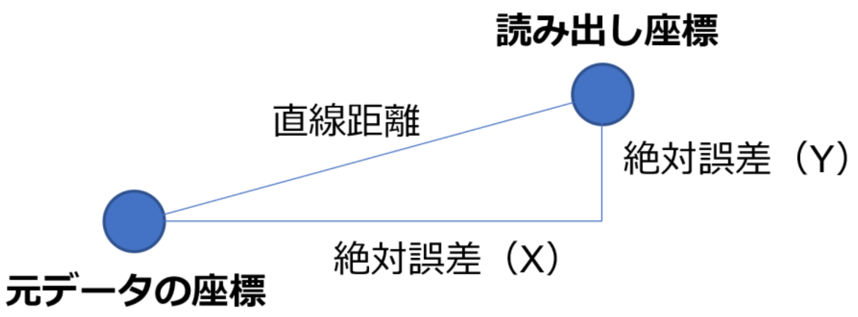


図 2：絶対誤差と直線距離

考察

本論文では、グラフ画像内に保存された情報にアクセスし、グラフ画像のプロット位置をグラフ上の座標位置から自動で読み出すプログラムの作成方法を提供した。本論文の方法は、ベクター画像の中でも EMF 画像を対象としたが、他のベクター画像フォーマットでも内部情報にアクセスできれば適用可能だと考えられる。一方で、ラスター画像フォーマットでは内部情報が 1 ピクセル単位での色の濃淡情報となるため、本論文の方法とは異なったパターン認識方法が必要となる。しかし、ラスター画像からベクター画像への変換を行うソフトウェアも存在するため、変換後に本方法を適用することもできる。ただし、変換によってすべての情報が完全に保存されるわけではない点には注意が必要である。

EMF 画像からのレコード情報の読み出しのために C++プログラムを作成したが、SAS でも同様の処理を行うことは可能だと思われる。SAS でもバイナリデータの読み込みは可能であり、内部の文字列情報は unicode 関数によって変換できる。

本論文では解析対象を散布図とし、プロット点は塗りつぶした黒円としたが、他の色やマーカーの種類でも適応でき、プロット点の種類ごとに読み出すこともできる。これは元の EMF 画像に保存されているレコード情報にアクセスするだけで済む。また、読み出し対象が直線や曲線であっても適応でき、連続的なプロット位置の情報を読み出すことができる。

本論文では散布図を例としたが、異なるグラフの種類でも適用可能だと考えられる。描画オブジェクトの内、X 軸 Y 軸と目盛を識別したが、同様にして、グラフの種類の識別も可能である。例えば、箱ひげ図であれば、特徴的な四角と上下に伸びるヒゲのパターンを識別できるように設計可能であると考えられる。また、カプラン・マイヤー曲線であれば、階段状に折れ曲がった直線と打ち切り点のパターンが現れるはずである。それらのパターンを識別できるように事前に確率分布モデルを設計できれば、識別可能になる。

本論文で得られた結果は目視での読み出しより十分高精度であった。さらに偶然ではあるが、今回作成した EMF 画像のレコード情報から X 軸と Y 軸は 2 度重ねて描画されていることが分かった。目視ではこれらの 2 度書かれた直線を区別することは全くできないが、目視で区別ができないような重なり合った 2 つのプロットを EMF 画像のレコード情報にアクセスすることで、区別ができるようになる可能性があるといえる。

上記のように本方法は本論文で解析対象とした散布図の EMF 画像以外にも適応できる可能性を秘めているが、認識したいパターンやデータ構造に合わせた適切な確率分布の仮定が必要になる。本論文で識別する描画オブジェクトのクラスは表 1 : 識別するレコードの分類に記載したものであるが、グラフの種類の識別にはさらに確率分布を仮定する必要がある。また、本論文ではサンプル画像として Proc SGPLOT を用いて散布図を作成したが、Proc GPLOT を用いて見た目上同じグラフを作成したとしても、内部のレコードは異なることが分かっている。そのため、グラフを作成する方法に合わせたレコード情報の前処理が必要になる。上記のような事前の確率分布の仮定を緩和するために機械学習のような方法で確率分布の構造を学習させることも可能であるが、十分な学習データを集める必要がある。

結論

臨床試験において、解析結果がグラフ画像である解析の品質管理をダブルプログラミングによって行う場合、結果の照合作業は人による目視作業となる場合が多く、作業に時間がかかり、その品質は担当者によって変動しうる。本論文では EMF 画像の内部レコード情報にアクセスし、パターン認識を行う方法を提供した。これによって、照合作業を自動化可能であり、短時間に一定の精度で照合作業を行うことができる。本論文での解析対象は Proc SGPLOT で作成した散布図の EMF 画像としたが、他のグラフの種類や画像の種類にも適応可能である。適切な確率分布の仮定を行うことで解析対象を拡張可能であり、臨床試験で頻繁に作成されるグラフに対する解析方法を確立することで、作業の効率化を達成できる。

本論文は業務効率化にベイズパターン認識を応用したものであり、他の業務でも適用可能なものはあると考えられる。グラフ画像が与えられたときの、X 軸や目盛、グラフ種類の識別などは人が見れば問題なく実行できる。グラフ画像 1 枚程度であれば問題なく処理できるが、臨床試験の場合は数 100 枚に及ぶ場合もあり、人での作業は時間がかかってしまう。そのような場合にパターン認識を用いることで、自動化が可能であることを示した。本論文を通した事例共有によって、新たに業務効率化をするようなアイデアに繋がれば幸いである。

補足資料 1：確率分布

Four-parameter Beta Distribution

$$\mathrm{FBe}(X|\alpha,\beta,l,u) = \frac{(X-l)^{\alpha-1}(u-X)^{\beta-1}}{(u-l)^{\alpha+\beta-1}B(\alpha,\beta)}$$

ここで、$B(\alpha,\beta)$はベータ関数である。平均は$l+\left(\frac{\alpha}{\alpha+\beta}\right)(u-l)$、分散は$\frac{\alpha\beta(u-l)^2}{(\alpha+\beta)^2(\alpha+\beta+1)}$である。

最頻値は$\alpha>1$かつ$\beta>1$のとき$l+\left(\frac{\alpha-1}{\alpha+\beta-2}\right)(u-l)$となる。

デルタ分布

$$\mathrm{Del}(X) = \begin{cases} 1, X=0 \\ 0, X\neq 0 \end{cases}$$

カテゴリカル分布

$$\mathrm{Cat}(X|\boldsymbol{\pi}) = \prod_{m=1}^{M} \pi_m^{[X=m]}$$

ここで、[·]は Iverson bracket であり、内部の数式が真であるとき 1、偽なら 0 となる。

補足資料 2：定数パラメータの設定

表 4 は本論文中で用いた定数パラメータの値であり、$\boldsymbol{\theta}_2$は Four-parameter Beta Distribution のパラメータの組み合わせであり、$\boldsymbol{\pi}$はカテゴリカル分布のパラメータの組み合わせである。図 3、図 4、図 5 はそれぞれ Four-parameter Beta Distribution のパラメータの組み合わせで定まる確率密度関数を示している。図 3 は X 軸の始点の X 座標であり、グラフの原点に当たるため、EMF 画像内の左側に寄っていることを仮定している。図 4 は X 軸の始点の Y 座標であり、EMF 画像内の下側に寄っていることを仮定している。図 5 は X 軸の長さであり、EMF 画像の横幅の 8 割程度の長さとなることを仮定している。

表 4：定数パラメータの値

パラメータ		値
$\boldsymbol{\theta}_2$	α_1	10
	β_1	40
	l_1	0
	u_1	640
	α_2	10
	β_2	2.5
	l_2	0
	u_2	480
	α_3	10
	β_3	2.5
	l_3	0
	u_3	640
$\boldsymbol{\pi}$	π_1	1E-20
	$\pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7$	$(1-\pi_1)/6$

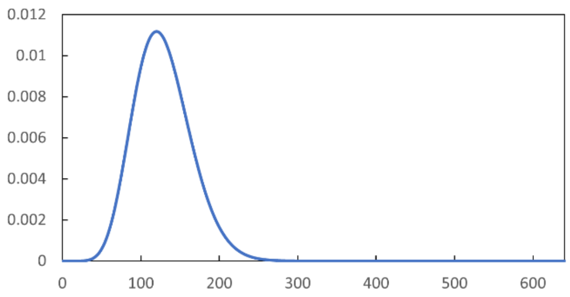


図 3：$\mathbf{FBe}(\boldsymbol{X_1}|\boldsymbol{\alpha_1}, \boldsymbol{\beta_1}, \boldsymbol{l_1}, \boldsymbol{u_1})$の確率密度関数

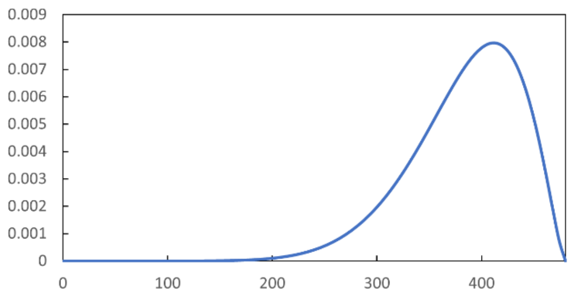


図 4：$\mathbf{FBe}(\boldsymbol{Y_1}|\boldsymbol{\alpha_2}, \boldsymbol{\beta_2}, \boldsymbol{l_2}, \boldsymbol{u_2})$の確率密度関数

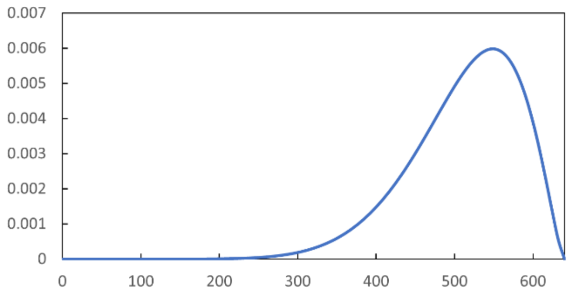


図 5：$\mathbf{FBe}(\boldsymbol{X_2 - X_1 | \alpha_3, \beta_3, l_3, u_3})$の確率密度関数

参考文献

[1] “[MS-EMFPLUS]: Enhanced Metafile Format Plus Extensions,” 2023 年 8 月. [オンライン]. Available: https://learn.microsoft.com/ja-jp/openspecs/windows_protocols/ms-emfplus/5f92c789-64f2-46b5-9ed4-15a9bb0946c6.

[2] “RStan: the R interface to Stan,” 2023 年 8 月. [オンライン]. Available: https://mc-stan.org/users/interfaces/rstan.

How to use the GAMs for Big Data.

ビッグデータに対する GAM の活用の仕方

古川敏仁

近年、データの集積が進み、いわゆる Big-Data の活用が盛んになっている。しかしながらデータの特質を良く捉えているとは言い難い解析結果も多いのが現実である。そこで今回は、宮田眼科病院で収集された眼から検出された 8642 例の菌データ（宮田 和典、岩崎琢也らによる）を用いて、GAM(generalized additive model)という手法を用いればその特質を良く捉える一助になることを、具体例をもとに説明する。

眼から検出される同一タイプの菌は、通常、1 眼 1 株であるが、まれに 2、3・・と複数検出される場合もある。検出されない場合は 0 株であるから、菌株の分布は 0,1,2/・・の典型的な Poisson 分布となる(Table1)。今、ある変数、例えば年齢と検出される菌株数の関係を考えるとする。このような場合、従来は GENMOD や GLIMMIX を用いた Poisson 回帰モデル（いわゆる一般化線形モデル）を用いて解析するのが一般的であった。Poisson 回帰モデルは応答 Y と説明変数 X との間に

Log(y)=log(c+a x) c:intertceot a:regression coefficients

という関係を仮定したモデルである。今回、応答を菌数、年齢を説明変数とすれば、log(菌数)と log（年齢）との間には直線関係を仮定して解析することになる。従来のデータ数がそれほど多くない時代はこれでもしょうがなかったが、近年、データの集積が進むとこの直線性の仮定を確かめる明確な、身近な利用可能な手段が発展してきた。その一つが GAM である。GAM スプライン関数を使えば説明変数上の応答変数の平均値の平滑化された軌跡を描くことができる。すると、Fig1 に示すように平均値の説明変数上の軌跡が直線なのか、直線近似はしない方が良いのかが判断できる。たとえば、図 1①が log linear model における説明変数 X と応答 Y の直線関係を示すものであるが、現実には直線ではなく、②のように X の途中から Y の変化が始まったり、極端な場合は③のように U 字型に変化する場合もある。実際に A 菌種について菌数の年齢上の平均値の軌道と 95%信頼区間を GAM で求めてみると Fig2 のようになった(PGM1 参照)。｛ただし、GAM の場合、その特質上から平均値の 95%信頼区間は求めることはできないので、復元抽出に基づく bootstrap 法で求めている。bootstrap 法は測定条件、施設、年次などの条件をどのように固定するか注意が必要な場合も多いが、ここでは、固定条件のないもっとも単純な例で示す｝例えばこれに従来のように単純な log liner model（Poisson 回帰モデル）をあてはめてみると、Fig3、Table3 のような結果となった。菌数は年齢が 10 歳増加するごとに 1.12（1.04~1.20）倍増化して、p=0.0027 である。果たしてこの結論で良いのだろうか？問題は年齢が 20～100 歳まで、いずれの時点でも均一に増加するという仮定を置くことである。Fig3 で確認しよう。モデル推定値は黒線のように 20～100 歳まで均一に増加しているが、実際の GAM 推定値はそうではない。70 歳ぐらいまで菌数はほぼ変わらず、70 歳ぐらいか

ら急に増加しているように見える。また、モデル推定値の 95%信頼区間からも GAM 平均推定値は 47 歳以下、86 歳以上で外れている。つまり、少なくともこの範囲は、線形 Poisson モデルは適切に菌数を予測していないことになる。もう少し Poisson モデルの推定精度を上げるためには、このようなシンプルな曲線では説明変数に関して、1 次（直線）の代わりに 2 次、3 次曲線をあてはめれば、格段に推定精度は向上する。さらに、得た関数の微分曲線とその 95%信頼区間を求めれば、変曲点はどこなのか、まだどの時点から確からしく増加するのか、減少するのかが分かり、非常に数多くの知見が得られる。しかし、この微分曲線は数学的な知識を持つ者にとっては非常に分かりやすく、良い解析を行ったと思えるのだが、数学にあまり馴染みのない者たちにとっては直感的に分かりにくいということもある。そこで、今回は、特定区間を複数の直線（1 次式）に置き換えた、直感的に分かりやすい方法を例示する。

それは、直線（線形）多項式を使う方法である。GAM 曲線(Fig2)を見ると年齢 70 歳ぐらいから菌数が増加し始めているように見える。そこで以下の 2 つの多項式を年齢 X に関して作成することとする(PGM2 参照)。

　if x<70 then X1=70-x

　if x≧70 then X2=x-70　　(0)

こうすることで、X1 と X2 は 70 で接合する 2 つの直線となる。

　一般的な X 年齢から Y 菌数を予測する一次回帰式は Y=a0+b0 X　(1)となるが

2 公式直線の場合はる回帰式は以下となる。

　Y=a1+b1 X1+b2 X2　(2)

ここで a0 は切片（x=0）の値であるが、(2)式の場合は x=70 (接合点)の値である。

一般的な回帰式では、a0 は X（Y に関しても）重心の近くにある方が b0 との相関が少なく、b0 の推定精度は良くなるが、a1 に関しても同様である。

　Poisson モデルに(0)の一次 2 項式をあてはめた結果を Fig4、Table4 に示す。Fig4 では、いずれの時点においても、Poisson モデルから推定した菌数は GAM から推定した菌数 95%信頼区間の中にあり、また、逆に、GAM から推定した菌数は Poisson モデルから推定した菌数の 95%信頼区間の中にある。つまり、Poisson モデルの一次 2 項式はかなり現状を良く表しているといえる。一次 2 項式モデルからは、測定された年齢 20~70 の間、菌数はほぼ変わらず、70 を過ぎると菌数は、平均　10 歳ごとに 1.39(1.22~1.59)倍ずつ増加することが分かる(p= <.0001)。多重性の調整を事前にすると決定していたわけではないので、p 値の有意性は問えないが、この場合、p 値は参考にある数値である。

　このように、シンプルな一次 2 項式からは、医学的な直観から求まる仮説、つまり、年齢 20~70 の間、菌数はほぼ変わらず、70 を過ぎると菌数は急激に増加し始める。を統計 k 的に裏付け記述することができ、また、参考とする p 値も求めることができる。

　Fig5 は、B 菌の年齢-菌数の関係を GAM（赤線）で示したものである。B 菌において菌数は、77 歳ぐらいにピークがある U 字型の分布をしている。これに従来の linear Poisson

モデルをあてはめると。観測者数の多い 80 歳以上は実際御分布に近い GAM では、歯数は減少しているのに対し、従来の linear Poisson モデルでは、全体を通して単調増加となるため、増加と解析されてしまう、明らかな誤解析である。しかも 80 歳以上は例数も多く、また、臨床的にも興味ある領域である、そこがこのような誤解析では、正しい医学的な判断はできない。そこで、今回は年齢 X に対し、以下のような一次 2 項式をあてはめてみた。

if x<73 then X1=73-x ①
if x≧73 and x<82 then X1=0 and X2=0 ②
if x≧82 then X2=x-82 ③ (3)

ちなみに、このモデル②の部分は、ピークの平坦な部分を示すものであり、ピークの値は Intercept として求めることができる。

結果は Fig6 のようであり、20 歳から 77,78 ぐらいまでは歯数は増加し、80 歳以上は歯数が減少する様が捉えられている。

ちなみに、Table6 からは、20~73 歳までは、年齢が 10 歳変化するごとに 1.12(1.08~1.16)倍増化し、83 歳以上は、0.76(0.64~0.90)倍ずつ減少すると予想された（いずれも参照 p 値<0.05）。

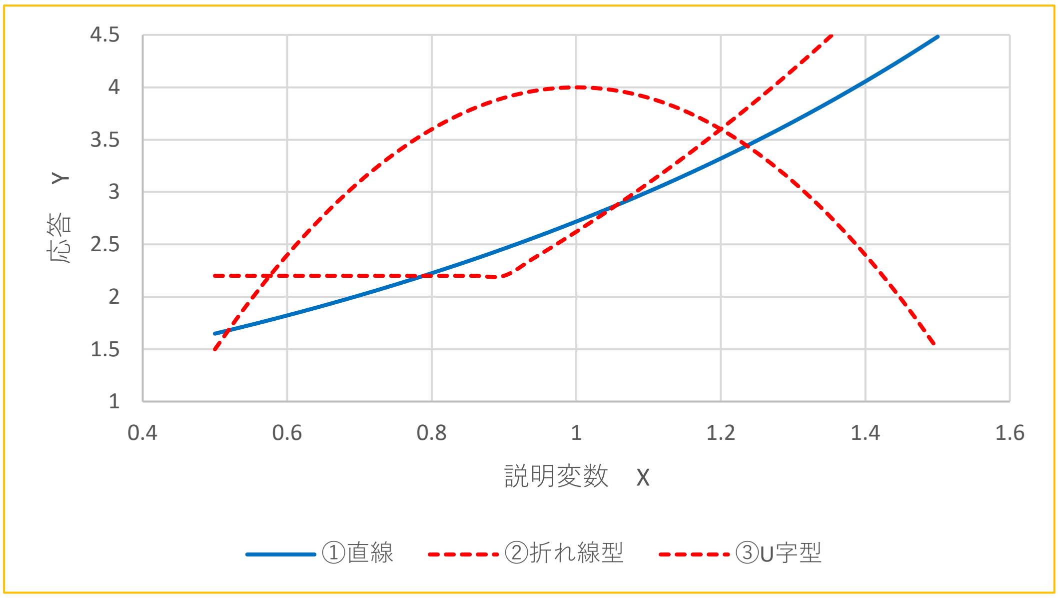
まとめ　大規模データになると、従来では評価不可能であった局所的な変化や関係が解析できるようになり、従来では得られなかった新たな臨床的知見が得られるようになった。そこで活躍する手法の一つが GAM である。GAM は曲主的変化を視覚的に得られるようになるばかりでなく、局所的な他の因子との関連を調整した、目的とする応答変数と説明変数の関係を明確にする優れた手法である。今回は、局所的変化を GAM を使い視覚的に、」また、2 項直線モデルを使い定量的に評価する手法を伝えた。次回は、局所的な他の因子との関連を調整した、目的とする応答変数と説明変数の関係を明確にすることを例示したい。

研究の続きとして、現在、ビッグデータを用いた調整という画題に取り組んでいる。
、ビッグデータでは、調査したい応答と特定項目の関係を、他の共変量を調整して行うということが統計モデルを使わなくてもできるようになる。それは共変量項目の共変量せっての特定の値を固定した条件で、応答と特定項目の関係を調べれば、それは、他の共変量の影響を除いた関係が調べられることになる。ただし、どんなに評価例数が増えても、共変量の項目数が増加したり、固定した共変量セットを複数調査しようとすれば、その共変量セットを固定化したもとで、の応答と特定項目のデータを十分集めるのには至難の業となる。そこで、名ならかの仮定がやはり必要となる。GAM は、他の一般化線形モデルとは違い、パラメータの加法性は仮定するが、趙県政は仮定しない。そのことで、より、リアルな共変量の影響を取り除くことができ、目的とする変数、共変量それぞれの、お互いの影響を調整した寄与の大きさを可視化できる。次回では、医学研究者の論文発表が終わり、そのデータ結果図表報告許可が下りたなら、その内容を報告したい。

Table1　一眼あたりの A 菌の分布

検出菌数	0	1	2	合計
眼数	8247	589	6	8842
割合(%)	93.27	6.66	0.07	

FIg1　応答 Y と説明変数 Y の関係

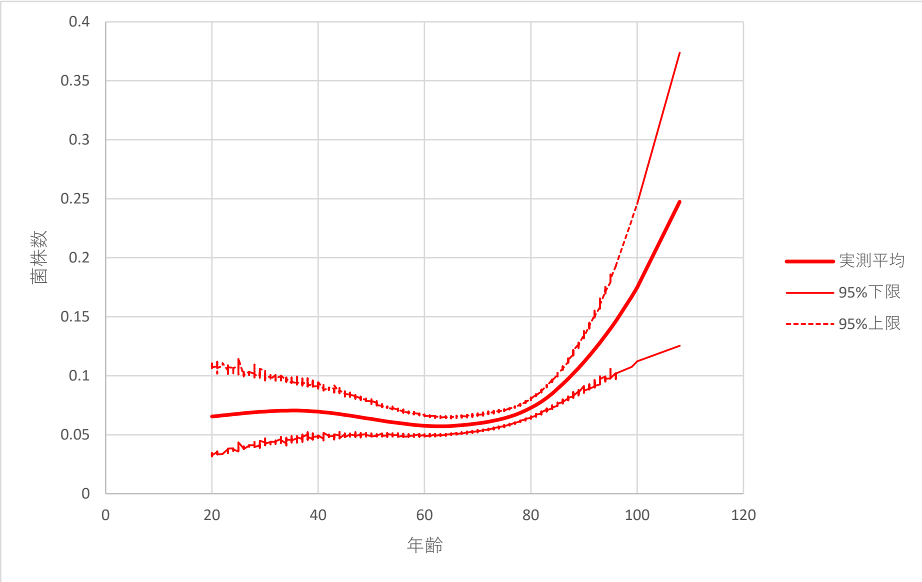


① は残像的に変化アするデータ

② X の屠龍時点から急速に増加する折れ線型変化

③ U 字型の変化

Fig3　　A 菌種について、GAM による年齢と菌数の関係の表示



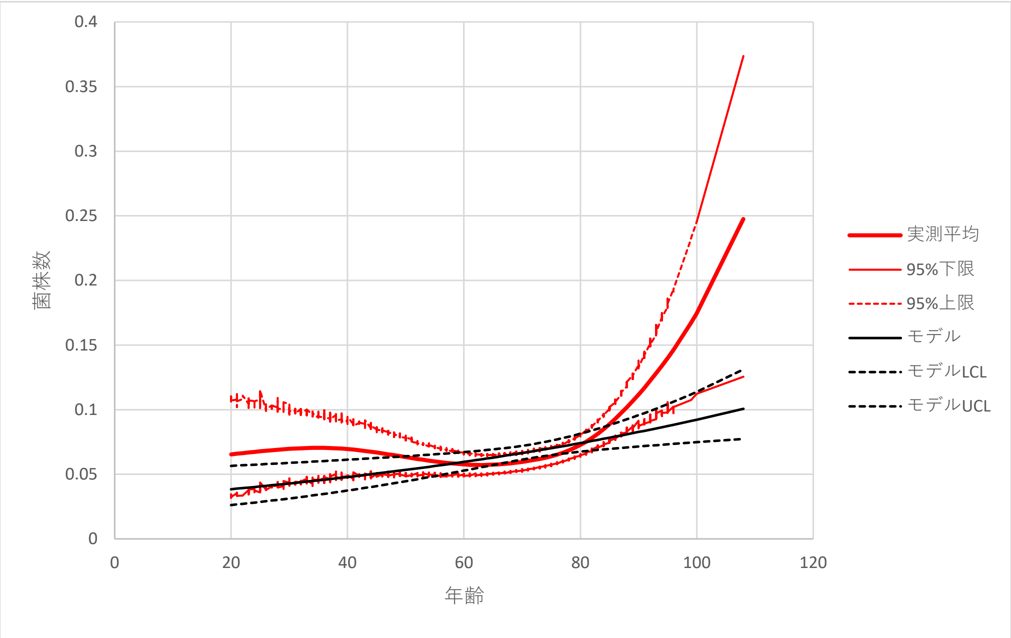
PGM1 参照

PGM1：GAM　Poisson 回帰による年齢-菌数曲線の推定と 95%信頼区間の導出

```
/*  年齢-菌数曲線の推定ルーチン***************/;
PROC GAM DATA=ANL ;;
  MODEL NUM   =   SPLINE(AGE, DF=3)
/ DIST=poisson;
  OUTPUT OUT=OUT1 all; RUN;
/**********入力データセット  ANL
NUM：一眼ごとの菌数、0,1,2
AGE：年齢
出力データセット
P_NUM：一眼ごとの推定菌数
AGE      ：年齢
/*************************        END***********/;

/*********95%信頼区間作成ルーチン   bootstrap 法  START************/;
proc surveyselect data=ANL   method=URS rep=1000
                 rate=1 seed=12345 out=ANL1;
  run;
```

```
  DATA ANL1;    SET ANL1;
DO K=1 TO NumberHits;    OUTPUT;        END;
RUN;

PROC GAM DATA=ANL1 ; BY Replicate;
  MODEL NUM    =    SPLINE(AGE, DF=3) / DIST=poisson;
  OUTPUT OUT=OUTF all; RUN;
  proc sort DATA=OUTF;BY NO RL AGE    Replicate    ;RUN;
proc univariate data = OUTF NOPRINT;
var P_num ; BY NO RL AGE    ;
output out = OUT2
            pctlpts    = 2.5 97.5                pctlpre = p ;      RUN ;
  /**********出力データセット  OUT2****************
  P2.5 P97.5   : 菌数の 95%信頼区間
AGE          : 年齢     ************/;
/*************************          END            ***********/;
```

Fig3　A 菌における年齢と菌数の関係　GAM と linear Poisson 回帰予測



赤線は GAM による予測

青線は linear Poisson 回帰予測

Table3　A 菌における年齢と菌数の関係　linear Poisson 回帰

年齢 10 歳単位の株数推定式（log linear model）

	母数推定値	95%下限	95%上限	p 値
Intercept	−3.4776	−4.0038	−2.9513	
AGE10	0.1095	0.0379	0.181	0.0027
尺度	0.9768	0.9768	0.9768	_

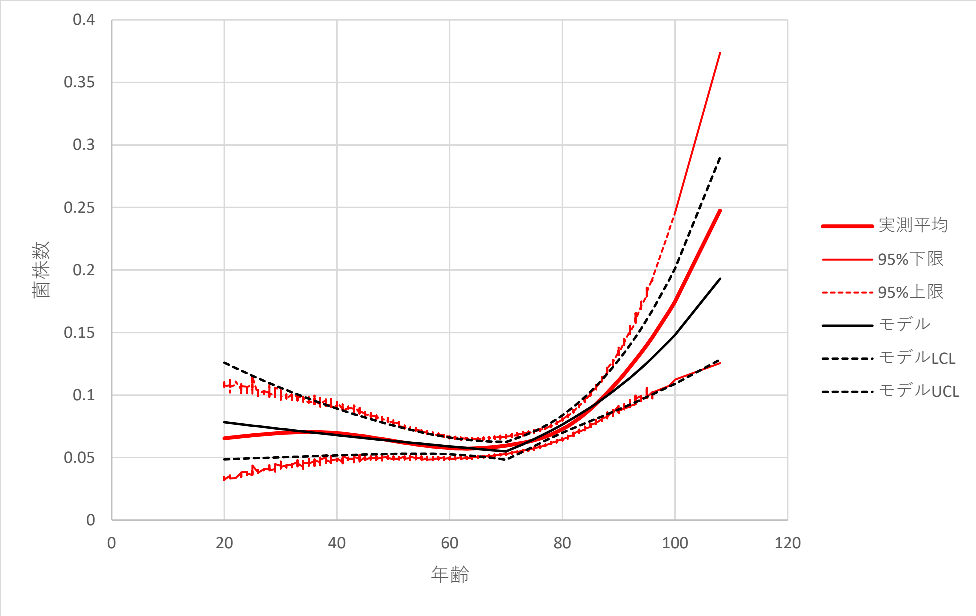
年齢 10 歳変化時の菌数変化比（年齢 20~100 歳）

	変化比	95%下限	95%上限	p 値
年齢 10 歳変化時の菌数変化比	1.12	1.04	1.20	0.0027

PGM2：直線 2 項式 (0,2)の作成

```
DATA ANL;SET ANL;
  IF AGE>=0 AND AGE<70 THEN DO;AGE10C1=(AGE-70)/10; END;
  IF AGE>=70    THEN DO;AGE10C2=(AGE-70)/10; END;
RUN;

/**** X1: AGE10C1 X2: AGE10C2******************/;
```

Fig4　　A 菌における年齢と菌数の関係　GAM と 2 項 linear Poisson 回帰予測



赤線は GAM による予測

青線は 2 項 linear Poisson 回帰予測

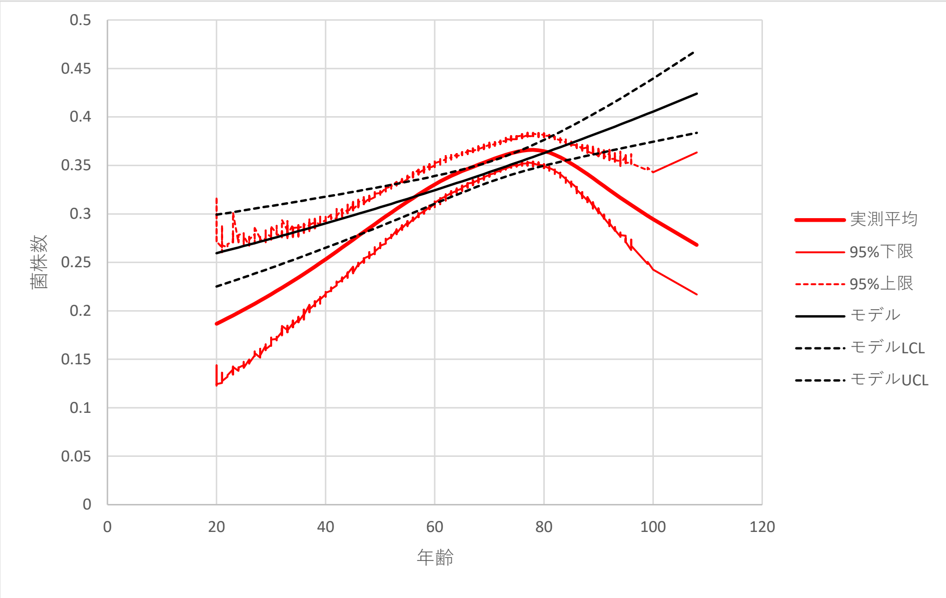
Table4　A 菌における年齢と菌数の関係　2 項 linear Poisson 回帰

年齢 10 歳単位の株数推定式（log linear model）

	母数推定値	95%下限	95%上限	p 値
Intercept	−2.9021	−3.0318	−2.7724	
年齢 70 歳未満(/10)	−0.0708	−0.178	0.0363	0.1949
年齢 71 歳以上(/10)	0.3308	0.2001	0.4615	<.0001
尺度	0.9769	0.9769	0.9769	_

年齢 10 歳変化時の菌数変化比	変化比	95%下限	95%上限	p 値
年齢 70 歳未満(/10)	0.93	0.84	1.04	0.1949
年齢 71 歳以上(/10)	1.39	1.22	1.59	<.0001

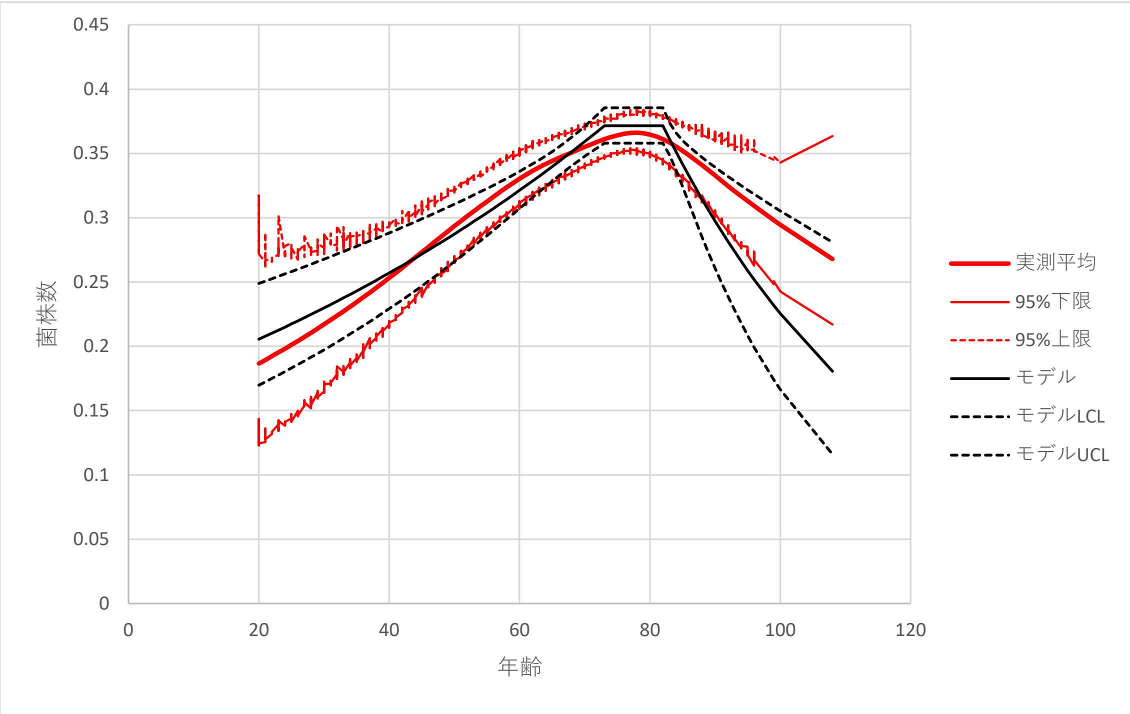
Fig5　　B 菌における年齢と菌数の関係　GAM と linear Poisson 回帰予測



赤線は GAM による予測

青線は linear Poisson 回帰予測

Fig6　　B 菌における年齢と菌数の関係　GAM と 2 項 linear Poisson 回帰予測



赤線は GAM による予測

青線は 2 項 linear Poisson 回帰予測

Table6　B 菌における年齢と菌数の関係　2 項 linear Poisson 回帰

年齢 10 歳単位の株数推定式（log linear model）

	母数推定値	95%下限	95%上限	p 値
Intercept	−0.9903	−1.0273	−0.9533	
年齢 73 歳以下(/10)	0.1116	0.0724	0.1508	<.0001
年齢 83 歳以上(/10)	−0.2776	−0.4524	−0.1028	0.0019
尺度	0.8496	0.8496	0.8496	_

年齢 10 歳変化時の菌数変化比	変化比	95%下限	95%上限	p 値
年齢 73 歳以下(/10)	1.12	1.08	1.16	<.0001
年齢 83 歳以上(/10)	0.76	0.64	0.90	0.0019

Hastie, T. J., and Tibshirani, R. J. (1990). Generalized Additive Models. New York: Chapman & Hall.

統計調査とSASによるサンプリング方法

○高田 浩成

（イーピーエス株式会社）

Statistical survey and sampling method with SAS

Hiroshige Takata
EPS Corporation

要旨

統計調査法は統計学の一分野として存在する。公的統計や民間統計が主な対象で、社会の実態を把握したり商業目的に使用したり、現代において重要な情報基盤となっている。その作成に携わる統計調査員の役割は企画・訪問・回収・監査・分析など様々で工夫が施されているが、本発表では、全数調査ではなく標本調査から全体像を推定する場合も多い中、コンピュータによるプログラミングが活用できる調査対象の標本抽出（サンプリング）に焦点を当て、SAS による実装方法を紹介する。SURVEYSELECT（調査／選択）プロシジャはそれを容易に実現でき、多くの機能を持つが、関連する情報は少ないように見受けられる。そこで、各標本抽出法から確率を用いたシミュレーションまでを説明・考察する。

キーワード：経済統計、proc surveyselect、標本抽出、無作為化、stratum、cluster、確率、反復試行

1. 背景

近年、官民を問わず合理的な証拠が重要視される時代となってきた。政府では政策の決定や施策の評価に、民間では経営戦略の立案や商品の企画・開発に多くのデータが活用されている。とくに、企業等が自ら調査によりデータを入手する動きがこれまで以上に活発化し、多用で大量のデータをいかに効率的に解析するかが経営・活動の成否を左右すると認識されるようになってきた。

統計調査の歴史として、古代から人口調査等は行われてきたが、封建的秩序が解体される中から近代的国家が形成される過程で、経済的視点からも国力についての関心が高まった背景がある。現代では、行政の運営を科学的に行うだけでなく、統計が民主主義社会に不可欠の情報を広く国民に提供する公共財であるとの役割の変化が認識されるようになり、グローバル経済や企業活動の情報インフラとして活用されるようになっている。

そして、より広い認識に立って、一定の条件（時間・空間・属性）で規定された集団に関して、記録された数値データの集まりを統計とする捉え方が一般的になっている。目的に沿って集計された結果にとどまらず、行政情報や観測データのように集計を目的とせずに記録されたデータに対しても、集団としての性質を表すものとして処理されると、統計が生成されることとなる。すなわち、統計が単なる数値データの集まりでは

ないのは、統計として作成する過程を経て、そこから何らかの意味ある情報を引き出せることにある。

2. 統計調査における標本設計

2.1 全数調査と標本調査

統計調査は、大きく、全数調査（悉皆調査）と標本調査の 2 つに区分される。全数調査は、調査の対象となっている集団（母集団）の統計単位を全て調べる調査であり、「本邦に居住している者」を対象と定めた国勢調査や、一部例外を除いて「全ての事業所・企業」を対象と定めた経済センサスなどが全数調査で実施されている。これに対して、標本調査は、母集団の一部を抽出して調査し、そこから母集団の状況を推定して結果を得るための調査である。

全数調査は、調査対象の全てである母集団を調査するので、ある時点における母集団の特性を正確に把握することが可能である。また、全体を細分化して、市町村や企業規模・産業小分類階級などの詳細な区分で集計することも可能になるなどのメリットがある。他方で、調査の実施や集計に要する費用や労力が多大となり、集計に時間が掛かるといったデメリットもある。

標本調査は全数調査と比較して、母集団の一部を標本として抽出して調査するので、全数調査と比べて、調査の費用・労力を削減でき、さらに、集計に要する時間も少なく、結果公表の早期化につながるというメリットがある。また、調査員の数が少なくて済むので、調査員の十分な訓練が可能となり、質の高い調査員の確保につながる。ただし、標本調査においては、母集団の一部を標本として抽出することによる誤差が生じるので、誤差の管理や、誤差を抑えるための推定の工夫が必要となる。

2.2 標本調査における標本抽出法

標本調査の方法は大きく、無作為抽出法(random sampling)と有意抽出法(purposive sampling)の 2 つの方法に区分できる。無作為抽出法は、母集団から標本を抽出する際に、抽出単位ごとに与えられる抽出確率を用いて標本を抽出する方法である。これに対し、有意抽出は、典型的・代表的と考えられる標本を主観的に抽出する方法である。

標本調査では、母集団の一部を標本として抽出して全体を推定することによる誤差が生じる。これを標準誤差といい、くじのあたりはずれに例えられる。無作為抽出では、抽出確率に基づき、標本誤差を調査結果から推定し、管理することが可能である。有意抽出では、抽出確率が明らかではないことから、標準誤差の測定や管理を行うことはできない。有意抽出は、試験調査などで特定の調査対象に対する調査方法の妥当性を確認したい場合などに用いられる。

種類の詳細については、SAS の実装とともに 4 章で紹介する。

3. SURVEYSELECT プロシジャの形式

標本抽出法は SAS の SURVEYSELECT プロシジャにより実装でき、無作為抽出法においては疑似乱数により実現することになる。SAS/STAT 12.1 からは RAND 関数と同様に Mersenne Twister と呼ばれるアルゴリズムに基づいて擬似乱数を生成される。このアルゴリズムは、旧来の RANUNI 関数などで採用されている乗算合同法に基づく方法と比較して、2^{19937}-1 という長い周期を取り、623 次元で均等に分布するため、統計的にも偏りがほとんど見られないと言える。

以下に SURVEYSELECT プロシジャの基本形を示す。原則として、抽出されたレコードは元の全ての変数を持つ。

```
proc surveyselect data=[dataset] out=[dataset] [option]；
   [statement]；
run；
```

代表的なオプション・ステートメントを表 1 及び表 2 に示す。

表 1： SURVEYSELECT プロシジャのオプション

オプション	別表記	役割	備考
sampsize=	n=	抽出数	レコード単位での指定
samprate=	rate=	抽出率	割合(0-1)または百分率(1-100)による表記
seed=		初期シード値	疑似乱数生成のため
reps=	rep=	反復回数	Replicate 変数の付与
outall		全データ出力	0：非抽出、1：抽出
outhits		重複レコードの別出力	復元抽出において使用
method=	m=	標本抽出法の指定	単純無作為抽出法(SRS)は省略可能
noprint		アウトプット画面の非出力	

表 2： SURVEYSELECT プロシジャのステートメント

ステートメント	別表記	役割	備考
id		出力変数の指定	
strata		層別化	事前 sort が必要 alloc=オプションの追加により標本配分法の実装
samplingunit	cluster	抽出単位の指定	クラスターの指定 PPS オプションの追加によりクラスター規模の考慮
size		重み付け	サンプリング方法によっては確率指定

抽出数(sampsize=)と抽出率(samprate=)はいずれかのオプションを用途に合わせて使用することになる。本発表では再現性の確保のため、初期シード値(seed=)は固定する。データを抽出するというより分割したい場合は outall オプションを使用することで、全レコードに対して 0 か 1 の数値型の Selected 変数(Selection Indicator)が付与され、2 分割された状態で結果のデータセットを出力できる。なお、実装方法により組み合わせられるオプション・ステートメントは異なることもあり、意図した標本抽出法を実施できているかどうかを確認できるように、結果ビューア・アウトプットの画面も表示することを推奨する。

4. 標本抽出法の実装

母集団が大きい場合には単純無作為抽出法よりも調査がしやすい抽出法を用いることがある。また、推定の精度をより高めるためにも、様々な抽出法が用いられることがある。それぞれにおいて、コード・出力データセット・アウトプット情報を示す。なお、本章では、便宜上簡略化した以下のテストデータに統一し、サンプリングを実施する。

表 3： テストデータ（データセット名：TEST）

	GROUP	NO
1	1	1
2	1	2
3	2	3
4	2	4
5	2	5

4.1 単純無作為抽出法(simple random sampling)

単純無作為抽出法は最も基本的な確率抽出法であり、どの抽出単位が抽出される確率も等しくなる。単純無作為抽出法をそのまま適用した場合、標本が特定の層に偏る可能性（推定結果が偏る可能性）と標本が広範囲に散らばる可能性（調査員の移動等の負担増加）の 2 つのデメリットがある。そこで、これらの課題に対応するために他の方法が適用されることがある。

なお、復元抽出法(sampling with replacement)では同じ要素が重複して標本に含まれる可能性があるのに対し、非復元抽出法(sampling without replacement)では重複することはない。そのため現実の標本抽出では非復元抽出法が用いられる。他方、復元抽出法のメリットは標準誤差等の計算が容易なことである。標本デザインによっては、複雑な非復元抽出法の計算式に代えて復元抽出法の計算式が用いられる。抽出率が小さい場合には、復元抽出でも要素が重複する可能性は低く、復元抽出法は非復元抽出法と同等とみなせる。

SURVEYSELECT プロシジャでは、単純無作為抽出法（非復元抽出法）は method=SRS になるが、省略可能である。

以下にコードと結果を示す。sampsize=3 として抽出数を指定した通り、ランダムで 3 レコード抽出されている。なお、復元抽出法を行う場合は、method=URS を指定し、さらに重複レコードも並べて表示する場合は outhits オプションが必要である。

```
proc surveyselect data=TEST out=OUT1 seed=1234 sampsize=3 ;
run ;
```

OUT1

	GROUP	NO
1	1	1
2	1	2
3	2	3

選択の方法	Simple Random Sampling

入力データセット	TEST
乱数シード	1234
標本サイズ	3
選択確率	0.6
サンプリングの重み	1.666667
出力データセット	OUT1

図 1： 単純無作為抽出法のコード及び結果

4.2 系統抽出法（等間隔抽出法）(systematic sampling)

現実の標本抽出場面では、単純無作為抽出法に代えて用いられることが多い標本デザインである。手順としては、大きさ N の母集団の各要素に 1 から N までの通し番号を付けてから、1 から N までの一様乱数 a を 1 つ発生させる。この a を開始番号(random start)という。そして、抽出間隔(sampling interval)d を定め、通し番号が a, a+d, a+2d, …, a+(n-1)d の要素を標本とする。ただし、要素の並び順に何らかの周期があり、それが抽出間隔に同期してしまうと、かえって標本誤差は拡大する点には注意する。

SURVEYSELECT プロシジャでは、method=SYS を指定し、その中に start=オプション（開始レコード）と interval=オプション（レコード間隔）を設定する。ただし、開始の数値は間隔以下にする必要がある。interval=オプションの代わりに samplesize=, samprate=オプションにより抽出数を指定して間隔を自動計算させることもできる（間隔を整数に計算できた場合に実行可能）。無作為抽出ではないため、seed=オプションは設定できない。

以下にコードと結果を示す。開始 2・間隔 2 で、データセットの上から下までその順に抽出されている（2 周目にあたる NO=1 のレコードは抽出されない）。

```
proc surveyselect data=TEST out=OUT2 method=SYS(start=2 interval=2)；
run；
```

選択の方法	Systematic Random Sampling

入力データセット	TEST
指定開始	2
Specified Interval	2
標本サイズ	2
選択確率	0.5
サンプリングの重み	2
出力データセット	OUT2

OUT2

	GROUP	NO
1	1	2
2	2	4

図 2：　系統抽出法のコード及び結果

4.3 層化抽出法（層別抽出法）(stratified sampling)

母集団をあらかじめ層(stratum)と呼ばれる複数の部分集団に分割しておき、どの層からも独立に所定の大きさの標本を抽出する方法である。例えば、都道府県や企業規模で調査対象を層化しておけば、どの都道府県や企業規模からも標本が抽出される。標本は母集団の縮図となり、単純無作為抽出法と比べて標本誤差が縮小すると期待できるため、層化抽出法は多くの調査で積極的に用いられる。標本デザインは層の間で異なることもある。

SURVEYSELECT プロシジャでは、strata ステートメントで層の変数を指定する。ただし、LIFETEST プロシジャ等と異なり、事前にソートが必要である。なお、strata ステートメントに alloc=オプションを付与することにより、標本配分法として、各層から抽出する標本の大きさを変えることもできる。比例配分法(alloc=PROPORTIONAL)では母集団において大きい層から大きい標本を抽出し、ネイマン配分法（最適配分法）(alloc=NEYMAN)ではそれに加えて散らばりの大きい層からも大きい標本を抽出する。

以下にコードと結果を示す。sampsize=オプションで抽出数の指定もできるが、samprate=オプションで抽出率を指定することで、層の規模により抽出数を変えている（抽出率を 50%としているが、3 レコードの層に対して、抽出数は繰り上げで 2 レコードとなっている）。

```
proc surveyselect data=TEST out=OUT3 seed=1234 samprate=0.5 ;
    strata GROUP ;
run ;
```

OUT3

	GROUP	NO	SelectionProb	SamplingWeight
1	1	2	0.5	2
2	2	4	0.6666666667	1.5
3	2	5	0.6666666667	1.5

選択の方法	Simple Random Sampling
層変数	GROUP

入力データセット	TEST
乱数シード	1234
層の標本抽出率	0.5
層の数	2
総標本サイズ	3
出力データセット	OUT3

図 3：　層化抽出法のコード及び結果

4.4 規模比例確率抽出法（確率比例抽出法）(probability proportional-to-size sampling)

層化抽出法において、規模で層化しネイマン配分を行うと、規模が大きい層ほど抽出率が大きくなり、推定量の標準誤差は縮小することになるが、確率比例抽出法（PPS と略される）は、規模で層化する代わりに、標本として選ばれる確率をその要素の規模に比例させ、より多くの大規模な要素を抽出する標本デザインである。

SURVEYSELECT プロシジャでは、method=PPS を指定し、size ステートメントにより変数に格納されている数値で重み付けする。なお、復元抽出には method=PPS_WR を指定し、系統抽出法とも組み合わせて method=PPS_SYS を指定することもある。

以下にコードと結果を示す。格納数値の大きい GROUP=2 を優先して抽出されている。

```
proc surveyselect data=TEST out=OUT4 method=PPS seed=1234 sampsize=4 ;
    size GROUP ;
run ;
```

OUT4

	NO	GROUP	SelectionProb	SamplingWeight
1	2	1	0.5	2
2	3	2	1	1
3	4	2	1	1
4	5	2	1	1

選択の方法	PPS, Without Replacement
サイズ測定	GROUP

入力データセット	TEST
乱数シード	1234
標本サイズ	4
出力データセット	OUT4

図 4：　PPS のコード及び結果

4.5 多段抽出法(multistage sampling)

多段抽出法の概念として、集落抽出法（クラスター抽出法）が存在する。母集団を分割した集落(cluster)を抽出単位として抽出を行い、選ばれた集落内の全ての要素を標本とする方法である。集落抽出法が用いられるのは、調査対象である要素のリストは入手が困難だが、集落のリストは利用可能な場合である。また、調査対象が集落内に寄り集まっているため、面接調査のコストを削減できるというメリットもある。標本世帯

が全国に点在するよりは、いくつかの地域に集中している方が効率的に訪問できる。一般に、集落抽出法は単純無作為抽出法よりも標本誤差が大きいが、その拡大を抑える方法はいくつかある。集落を自由に構成できるのであれば、各集落は小さくするとともに、集落内には異質な要素を含めると良い。例えば、エリアサンプリングとして、地図上の対象地域を分割した小地域を抽出単位とし、選ばれた小地域内の世帯数を標本とする方法がある。

集落抽出法において、選ばれた集落の中でさらに一部の要素だけを抽出すると、二段抽出法となる。その次は三段抽出法というように、一般に、各段で選ばれた集落の中で、さらに抽出を繰り返す方法を多段抽出法という。

SURVEYSELECT プロシジャでは、samplingunit(cluster)ステートメントにより抽出単位を設定する。

以下にコードと結果を示す。sampsize=1 の抽出数を指定することで GROUP=2 の 1 グループの全レコードが抽出された。

```
proc surveyselect data=TEST out=OUT5 seed=1234 sampsize=1 ;
    cluster GROUP ;
run ;
```

OUT5

	GROUP	NO
1	2	3
2	2	4
3	2	5

選択の方法	Simple Random Sampling
抽出単位変数	GROUP

入力データセット	TEST
乱数シード	1234
標本サイズ	1
選択確率	0.5
サンプリングの重み	2
出力データセット	OUT5

図 5： クラスター抽出法のコード及び結果

5. その他のサンプリングの実装

SURVEYSELECT プロシジャは統計調査における標本抽出法以外にも様々な機能を持ち、method=オプション等で設定できるため、その一部を紹介する。Bernoulli Sampling, Poisson Sampling では、標本が抽出される確率（前章までの抽出率とは扱いは異なる）を指定するが、実際の抽出数には変動が生じることにもなる。確率の精度については十分な回数の反復試行により確認も行う。本章では、以下のテストデータに対してサンプリングを実施する。

表 4： テストデータ（データセット名：TEST2）

	NO	Pi
1	1	0.2
2	2	0.4
3	3	0.6
4	4	0.8
5	5	1

5.1 Bernoulli Sampling

SURVEYSELECT プロシジャの method=BERNOULLI では、samprate=オプションにより、1 つの抽出確率を全レコードに適用する。母集団サイズを N、設定確率を P とすると、抽出数の期待値は平均 NP、分散 NP(1-P)となる。以下のコード・結果のように、いずれのレコードも 50%の確率で抽出されるようになっている。

```
proc surveyselect data=TEST2 out=OUTB method=BERNOULLI seed=1234 samprate=0.5 ;
run ;
```

OUTB

	NO	Pi
1	3	0.6
2	4	0.8

選択の方法	Bernoulli Sampling

入力データセット	TEST2
乱数シード	1234
選択確率	0.5
ユニットの総数	5
期待標本サイズ	2.5
標本サイズ	2
サンプリングの重み	2
調整済みサンプリングの重み	2.5
出力データセット	OUTB

図 6： Bernoulli Sampling のコード及び結果

100 万回反復試行(reps=1000000)を行うと、N = 5、P = 0.5 より、平均 2.5、分散 1.25 の期待抽出数に対し、実際に標本毎の集計と反復回数毎の平均・分散も以下の結果になり、想定された分布に収束する傾向が確認された。

	NO	COUNT	PERCENT
1	1	499720	49.972
2	2	500135	50.0135
3	3	500319	50.0319
4	4	499561	49.9561
5	5	500373	50.0373

	FREQ	MEAN	VARIANCE
1	1000000	2.500108	1.2513632397

図 7： Bernoulli Sampling 反復試行による集計及び平均・分散の結果

5.2 Poisson Sampling

SURVEYSELECT プロシジャの method=POISSON では、size ステートメントで指定した変数に格納された数値を抽出確率として各レコードに適用する。母集団サイズを N、設定確率を Pi とすると、抽出数の期待値は平均 ΣP_i、分散 $\Sigma P_i(1-P_i)$となる。以下のコード・結果のように、適用された確率の Pi 変数が SelectionProb 変数(Probability of Selection)に変わり、格納されていた確率が大きいレコードほど抽出されやすくなっている。

```
proc surveyselect data=TEST2 out=OUTP method=POISSON seed=1234 ;
    size Pi ;
run ;
```

選択の方法	Poisson Sampling
選択確率	Pi

入力データセット	TEST2
乱数シード	1234
ユニットの総数	5
期待標本サイズ	3
標本サイズ	3
出力データセット	OUTP

OUTP

	NO	SelectionProb	SamplingWeight
1	3	0.6	1.6666666667
2	4	0.8	1.25
3	5	1	1

図 8： Poisson Sampling のコード及び結果

100 万回反復試行(reps=1000000)を行うと、N = 5、P_i = 0.2, 0.4, 0.6, 0.8, 1 より、平均 3、分散 0.8 の期待抽出数に対し、実際に標本毎の集計と反復回数毎の平均・分散も以下の結果になり、想定された分布に収束する傾向が確認された。

	NO	COUNT	PERCENT
1	1	200746	20.0746
2	2	400275	40.0275
3	3	600515	60.0515
4	4	800175	80.0175
5	5	1000000	100

	FREQ	MEAN	VARIANCE
1	1000000	3.001711	0.800500873

図 9： Poisson Sampling 反復試行による集計及び平均・分散の結果

5.3 Balanced Bootstrap Sampling

SURVEYSELECT プロシジャの method=BALBOOTSTRAP(BALBOOT)では、ブートストラップ法の一種として、均衡的な再標本化（リサンプリング）を行う。これにより、例えば、データが少ない時に活用し、標本のパターンを増やすことができる。以下のコード・結果のように、reps=オプションによる反復回数分だけレコードを複製し、Replicate (Sample Replicate Number)変数に反復回数にあたるグループ連番が格納され、各グループに標本を配分し直している（グループ内で重複する標本も発生する）。

```
proc surveyselect data=TEST2 out=OUTBB method=BALBOOTSTRAP seed=1234 reps=3 ;
run ;
```

	Replicate	NO	P
1	1	1	0.2
2	1	2	0.4
3	1	2	0.4
4	1	4	0.8
5	1	5	1
6	2	1	0.2
7	2	3	0.6
8	2	3	0.6
9	2	4	0.8
10	2	5	1
11	3	1	0.2
12	3	2	0.4
13	3	3	0.6
14	3	4	0.8
15	3	5	1

OUTBB

選択の方法	Balanced Bootstrap

入力データセット	TEST2
乱数シード	1234
標本サイズ	5
繰り返し数	3
総標本サイズ	15
出力データセット	OUTBB

図 10： Balanced Bootstrap Sampling のコード及び結果

100 万回反復試行(reps=1000000)を行うと、以下の集計・統計量の結果になり、複製された全ての標本が均等に配分され、偏りのない分布であることが分かる。

	NO	COUNT	PERCENT	MEAN	VARIANCE	SD
1	1	1000000	100	500061.68	83346466255	288698
2	2	1000000	100	500260.18	83373352813	288744
3	3	1000000	100	499813.16	83463236468	288900
4	4	1000000	100	499974.22	83234634747	288504
5	5	1000000	100	499893.25	83249274583	288530

図 11： Balanced Bootstrap Sampling 反復試行による集計及び統計量の結果

6. 結語

統計調査の重要性とそれにおけるサンプリングの役割から SAS による実装方法を紹介してきた。SURVEYSELECT プロシジャは、標本調査法を実装するにあたり、様々な方法に対応可能で、非常に有用であることを確認できた。また、反復実行機能により、確率の再現性が高いことが確認され、確率的な面からはシミュレーションにも適していることも言える。今後、データの在り方が変化していく中で、本方法の意義を見直し、その複雑さを含めた機能も十分に利用し、調査分野のみならず、他の場面においても、サンプルデータ作成やアルゴリズム実装に活用する機会が増えれば幸いである。

参考文献

[1] SAS Institute Inc., SAS/STAT® 15.1 User's Guide - The SURVEYSELECT Procedure (2019)

[2] SAS Institute Inc., Statistical Business Analysis Using SAS9 (2018)

[3] 日本統計学会、経済統計の実際、東京図書(2022)

[4] 日本統計学会、調査の実施とデータの分析、東京図書(2023)

[5] 日本統計学会、統計学実践ワークブック、学術図書出版社(2020)

[6] 魚住龍史、浜田知久馬、RAND 関数による擬似乱数の生成、SAS ユーザー総会(2013)

SASによる粒子群最適化の実装

○折井悟

（イーピーエス株式会社）

Particle Swarm Optimization with SAS

Satoru Orii

(EPS Corporation)

要旨

粒子群最適化は，生物の群れのふるまいをモデルとした「群知能」と呼ばれる最適化アルゴリズムの一種である．粒子群最適化は他の最適化手法に比して複雑な処理を必要としないことが特徴であり，SAS の基本的な機能のみでも実装可能である．

本論文では，粒子群最適化の概要の解説および SAS での実装例の紹介を行う．

キーワード：Particle Swarm Optimization

粒子群最適化とは

粒子群最適化は，鳥や魚などの群れが餌を探す行動をモデルにした最適解探索アルゴリズムであり[1]，生物の群れのふるまいをモデルとして最適解を探索する「群知能」に分類される．粒子群最適化では，生物の個体に見立てた粒子を多次元空間内に多数配置し，各粒子がより良い解の情報を共有しながら空間内を移動していくことで最適解を探索する．

具体的な粒子群最適化の流れは以下のようになる．

1. 初期位置の設定

   解空間内にランダムな位置ベクトルと速度ベクトルを持った粒子 N 個を配置する．解空間の次元は評価関数における変数の数に等しい．

2. 速度の計算

   時刻 t における各粒子 i の位置ベクトル$\vec{x}_i$(t)および速度ベクトル$\vec{v}_i$(t)，各粒子がその時刻 t までに通った点のうち最も評価関数の値が高かった点(パーソナルベスト)の位置ベクトル$\vec{p}_i$(t)，粒子群全体がその時刻までに通った点のうち最も評価関数の値が高かった点(グローバルベスト)の位置ベクトル$\vec{g}$(t)を用いて，時刻 t+1 における各粒子の速度ベクトル$\vec{v}_i$(t+1)は次の式で計算される．

   $$\vec{v}_i(t+1) = w\vec{v}_i(t) + c_1r_1(\vec{p}_i(t) - \vec{x}_i(t)) + c_2r_2(\vec{g}(t) - \vec{x}_i(t))$$

   ここで，w, c_1, c_2 (>0)は事前に設定しておく重み，r_1, r_2 は 0 から 1 までの値を取る一様乱数である．

すなわち，時刻 t における速度ベクトル，時刻 t における位置からパーソナルベストへの方向ベクトル，時刻 t における位置からグローバルベストへの方向ベクトルの 3 つのベクトルの線形和により時刻 t+1 での速度が計算される．

3. 位置の更新

各粒子の時刻 t における位置ベクトル$\vec{x}_i$(t)と速度ベクトル$\vec{v}_i$(t+1)を足し合わせ，時刻 t+1 における各粒子の位置ベクトル$\vec{x}_i$(t+1)を求める．その後，$\vec{x}_i$(t+1)における評価関数の値を計算し，時刻 t におけるパーソナルベスト$\vec{p}_i$(t)またはグローバルベスト$\vec{g}$(t)における評価関数の値より大きい場合，それぞれ値を更新する．

2 および 3 の操作を 1 サイクルとして，設定した最大サイクル数に達するか，グローバルベストにおける評価関数の値が設定した閾値に達するまで繰り返す．最終的に最も評価関数の値が大きかった位置の情報を求められた解として出力する．

SAS での実装

Appendix 1 では例として以下に示す関数の最小値を求めるコードを示している．

$$f(x_1,\, x_2) = -(x_2+47)\sin\left(\sqrt{\left|x_2+\frac{x_1}{2}+47\right|}\right) - x_1\sin\left(\sqrt{|x_1-(x_2+47)|}\right) \quad (-512 \leqq x_1, x_2 \leqq 512)$$

この *f(x₁, x₂)*は最適化アルゴリズムの評価に用いられるベンチマーク関数である Eggholder Function[2]である．

以下，コード中の各マクロを解説する．

%PSO_execute は粒子群最適化の実行用のマクロで，問題に合わせた具体的な操作内容は後に示す各サブマクロに記述している．パラメータとしては最大サイクル数(max_time)，生成する粒子の数(n_ind)，速度の計算時に用いる重み(weight, acceleration_personal, acceleration_global)，評価関数の閾値(threshold)，評価関数の変数の数(varn)を与える．

データセット pso_0 では初期集団を生成している．**%init_create** は粒子に与えるランダムな初期位置および初期速度を生成するためのマクロである．例では変数 x_1, x_2 それぞれについて定義域の範囲で一様乱数を発生させて初期位置とし，定義域の大きさに適当な係数および一様乱数をかけて初期速度としている．

```
%macro init_create;
    array upper   upper1-upper&varn.;
    array lower   lower1-lower&varn.;
    array x        x1-x&varn.;
    array v        v1-v&varn.;
    array pb       pb1-pb&varn.;

    do i = 1 to &varn.;
        upper(i)=512;
```

```
        lower(i)=-512;
        x(i)=rand("uniform")*(upper(i)-lower(i))+lower(i);
        v(i)=0.5*(rand("uniform")-0.5)*(upper(i)-lower(i));
        pb(i)=x(i);
    end;
    %eval_function;
    fitness_pb=fitness;
%mend;
```

%eval_function は位置ベクトルから評価関数の値を求める際に用いるマクロである．例で扱っているのは $f(x_1, x_2)$ の最小値を求める問題であるため，$f(x_1, x_2)$ の値が小さいほど評価が高くなるよう，$f(x_1, x_2)$ の正負を逆にしたものを評価関数として用いている．

```
%macro eval_function;
    fitness=(x2+47)*sin(sqrt(abs(x2+x1/2+47)))+x1*sin(sqrt(abs(x1-(x2+47))));
%mend;
```

データセット gb では pso_0 から評価関数の値が最も高い位置をグローバルベストとして記録し，グローバルベストにおける評価関数の値をマクロ変数 fitness_gb に格納している．

以下の%do ループでは，グローバルベストにおける評価関数の値が閾値を上回っているかを判定し，上回っていればその時点のグローバルベストと評価関数の値を出力し実行を終了する．そうでなければ各粒子の情報を次の時刻の情報に更新するマクロ **%PSO_update** を実行する．閾値に達することなく設定した回数のループが終了した場合，グローバルベストと評価関数の値を出力し実行を終了する．

```
%do time=0 %to &max_time.;
    %if %sysevalf(&fitness_gb.>&threshold.,boolean) = 1 %then %do;
        data _null_;
            set gb(obs=1);
            array x         x1-x&varn.;
            array gb        gb1-gb&varn.;

            do i = 1 to &varn.;
                x(i)=gb(i);
            end;

            put "解は" (x:) (=) "  適合度は" fitness;
        run;
        %return;
    %end;
```

```
    %PSO_update(time=&time.,weight=&weight.,acc_p=&acceleration_personal.,acc_g=&acceleration_global.); /*t+1
秒の情報に更新*/
%end;
```

%PSO_update では時刻 t における情報をもとに，先程示した式を用いて時刻 t+1 における各粒子の位置と速度を計算している．以下の部分では，変数の値が定義域の外に出るような速度になった場合に定義域の端に位置を修正する処理をしている．

```
if x(i)>upper(i) then do;
    v(i)=v(i)-(x(i)-upper(i));
    x(i)=upper(i);
end;
if x(i)<lower(i) then do;
    v(i)=v(i)-(x(i)-lower(i));
    x(i)=lower(i);
end;
```

各粒子の位置から評価関数の値を計算した後，データセット gb に格納されたグローバルベストの評価関数の値と比較して上回っていればグローバルベストを更新する．

```
data gb;
    merge gb
            pso_%eval(&time.+1)(rename=(fitness=fitness_new) obs=1)
;
     array x          x1-x&varn.;
     array gb         gb1-gb&varn.;

     if fitness_new>fitness then do;
         do i = 1 to &varn.;
             gb(i)=x(i);
         end;
         fitness=fitness_new;
         call symput ("fitness_gb",cats(fitness));
         time_gb=&time.+1;
     end;

     keep time: gb: fitness;
run;
```

評価

Eggholder Function

作成したプログラムを実行し，アルゴリズムの評価を行った．

まず，前述の Eggholder Function の最小値を探索する設定でプログラムを 100 回実行した．
実行時の各パラメータは下記の通り設定した．
・最大サイクル数(max_time) : 100
・粒子数(n_ind) : 10000
・速度計算時の重み(weight, acceleration_personal, acceleration_global) : 0.9
・閾値(threshold) : 959.6406
・変数の数(varn) : 2

各実行において出力された解を評価関数の値の降順に並べ替えたものを Appendix 2 に示した．
Eggholder Function の$-512 \leq x_1, x_2 \leq 512$での真の最小値$f_{min}$は，$f(512, 404.2319) = -959.6407$ であることが知られている．今回の実行では，100 回中 97 回の実行で大域的最適解の近傍へ到達した．また，実行に要した時間は 1 回あたり 7.75 秒であった．
この結果は今回実装したアルゴリズムが実用的な時間で十分な探索を実行できることを示すものと考える．
また，以前実装した遺伝的アルゴリズムで同様の最適化を実施した際は 100 回の実行中 38 回で大域的最適解の近傍へ到達し，実行時間は 1 回あたり 45.55 秒であった．このことから，ある種の最適化問題においては今回実装したアルゴリズムは遺伝的アルゴリズムより高速・高確率で大域的最適解へ到達することができるといえる．

まとめ

本稿では，Base SAS，SAS/STAT の機能を用いて粒子群最適化を実装し，ベンチマーク関数の大域的最適解を探索することによりアルゴリズムの評価を実施した．今回評価に用いたベンチマーク関数では，実用的な時間で大域的最適解に到達できることが示された．
今後の課題として，他のベンチマーク関数での評価，実業務での応用可能性の検討等を行いたいと考えている．

参考文献

[1] Kennedy, J. & Eberhart, R. “Particle Swarm Opitimization”, Proceedings of ICNN'95 - International Conference on Neural Networks, pp.1942-1948, 1995.
[2] Adorio, E. P., & Diliman, U. P. “MVF - Multivariate Test Functions Library in C for Unconstrained Global Optimization”, Jan. 14, 2005
http://www.geocities.ws/eadorio/mvf.pdf (Accessed Aug 11, 2023)

Appendix 1

```
%macro PSO_execute(max_time=,n_ind=,weight=,acceleration_personal=,acceleration_global=,threshold=,varn=);
    data pso_0;
      time=0;
      do _= 1 to &n_ind.;
```

```
        %init_create;
        output;
    end;

    keep time upper: lower: x: v: pb: fitness:;
run;

proc sort data=pso_0;
    by desending fitness;
run;

data gb;
    set pso_0(obs=1);
    array x         x1-x&varn.;
    array gb        gb1-gb&varn.;

    do i = 1 to &varn.;
        gb(i)=x(i);
    end;

    call symput ("fitness_gb",cats(fitness));

    keep time gb: fitness;
run;


%do time=0 %to &max_time.;
    %if %sysevalf(&fitness_gb.>&threshold.,boolean) = 1 %then %do;
        data _null_;
            set gb(obs=1);
            array x         x1-x&varn.;
            array gb        gb1-gb&varn.;

            do i = 1 to &varn.;
                x(i)=gb(i);
            end;

            put "解は" (x:) (=) "  適合度は" fitness;
        run;
```

```
            proc datasets lib=work memtype=data nolist;
                delete
                pso_0-pso_%eval(&max_time.+1)
                result;
            quit;
           %return;
       %end;

        %PSO_update(time=&time.,weight=&weight.,acc_p=&acceleration_personal.,acc_g=&acceleration_global.);
/*t+1 秒の情報に更新*/
    %end;



    data _null_;
        set gb(obs=1);
        array x         x1-x&varn.;
        array gb        gb1-gb&varn.;

        do i = 1 to &varn.;
            x(i)=gb(i);
        end;

        put "解は" (x:) (=) "  適合度は" fitness;
    run;

    proc datasets lib=work memtype=data nolist;
        delete
        pso_0-pso_%eval(&max_time.+1)
        result;
    quit;
%mend;

%macro eval_function;
   fitness=(x2+47)*sin(sqrt(abs(x2+x1/2+47)))+x1*sin(sqrt(abs(x1-(x2+47))));
%mend;

%macro init_create;
   array upper    upper1-upper&varn.;
```

```
    array lower    lower1-lower&varn.;
    array x        x1-x&varn.;
    array v        v1-v&varn.;
    array pb       pb1-pb&varn.;

    do i = 1 to &varn.;
        upper(i)=30;
        lower(i)=-30;
        x(i)=rand("uniform")*(upper(i)-lower(i))+lower(i);
        v(i)=0.5*(rand("uniform")-0.5)*(upper(i)-lower(i));
        pb(i)=x(i);
    end;
    %eval_function;
    fitness_pb=fitness;
%mend;

%macro PSO_update(time=,weight=,acc_p=,acc_g=);
    data pso_%eval(&time.+1);
        merge pso_&time.
                gb(keep=gb: time);
        by time;
        time=&time.+1;

        array upper    upper1-upper&varn.;
        array lower    lower1-lower&varn.;
        array x        x1-x&varn.;
        array v        v1-v&varn.;
        array pb       pb1-pb&varn.;
        array gb       gb1-gb&varn.;

        rc1=rand("uniform");
        rc2=rand("uniform");

        do i=1 to &varn.;
            v(i)=v(i)*&weight.+&acc_p.*rc1*(pb(i)-x(i))+&acc_g.*rc2*(gb(i)-x(i));
            x(i)=x(i)+v(i);
            if x(i)>upper(i) then do;
                v(i)=v(i)-(x(i)-upper(i));
                x(i)=upper(i);
```

```
            end;
            if x(i)<lower(i) then do;
                v(i)=v(i)-(x(i)-lower(i));
                x(i)=lower(i);
            end;
        end;

        %eval_function;
        if fitness>fitness_pb then do;
          do i = 1 to &varn.;
              pb(i)=x(i);
          end;
          fitness_pb=fitness;
        end;
run;

proc sort data=pso_%eval(&time.+1);
        by desending fitness;
run;

data gb;
        merge gb
                pso_%eval(&time.+1)(rename=(fitness=fitness_new) obs=1)
;

        array x         x1-x&varn.;
        array gb        gb1-gb&varn.;

        if fitness_new>fitness then do;
            do i = 1 to &varn.;
                gb(i)=x(i);
            end;
            fitness=fitness_new;
            call symput ("fitness_gb",cats(fitness));
            time_gb=&time.+1;
        end;

        keep time: gb: fitness;
run;
```

```
%mend;
```

Appendix 2

x_1	x_2	$f(x_1, x_2)$
512	404.2317961	-959.6406627
512	404.2319042	-959.6406627
512	404.2319175	-959.6406627
512	404.2315827	-959.6406627
512	404.2315379	-959.6406626
512	404.2312947	-959.6406624
512	404.2323307	-959.6406624
512	404.2312068	-959.6406623
512	404.2324064	-959.6406623
512	404.2324087	-959.6406623
512	404.2325454	-959.6406621
512	404.2325643	-959.6406621
512	404.2326486	-959.6406619
512	404.2327235	-959.6406618
512	404.2327328	-959.6406617
512	404.2327457	-959.6406617
512	404.230861	-959.6406617
512	404.2328934	-959.6406614
512	404.2305914	-959.640661
512	404.2330395	-959.640661
512	404.2331489	-959.6406607
512	404.2332861	-959.6406602
512	404.2301291	-959.6406595
512	404.2334866	-959.6406595
512	404.2300933	-959.6406594
512	404.2335196	-959.6406594
512	404.2300544	-959.6406592
512	404.2336201	-959.640659
512	404.2298798	-959.6406585
512	404.2298294	-959.6406583
512	404.2297447	-959.6406579
512	404.2296695	-959.6406575
512	404.229594	-959.6406572
512	404.2295738	-959.6406571
512	404.2340665	-959.6406569
512	404.2340984	-959.6406567
512	404.2341305	-959.6406566
512	404.2342	-959.6406562
512	404.2342203	-959.6406561
512	404.22925	-959.6406553
512	404.2344082	-959.640655
512	404.234551	-959.6406541
512	404.234622	-959.6406537
512	404.2289036	-959.6406531
512	404.2288904	-959.6406531
512	404.2286653	-959.6406515
512	404.2351158	-959.6406503
512	404.2352544	-959.6406492
512	404.2281308	-959.6406474
512	404.2281192	-959.6406473
512	404.2356678	-959.6406458
512	404.227928	-959.6406456
512	404.2279003	-959.6406454
512	404.235745	-959.6406451
512	404.2359535	-959.6406432
512	404.2276444	-959.640643
512	404.2273658	-959.6406403
512	404.2362713	-959.64064
512	404.236369	-959.640639
512	404.2272064	-959.6406387
512	404.2272026	-959.6406386
512	404.236428	-959.6406384
512	404.2270043	-959.6406365
512	404.2366977	-959.6406355
512	404.2269023	-959.6406354

512	404.2268605	-959.6406349
512	404.2267748	-959.640634
512	404.2267266	-959.6406334
512	404.2369195	-959.640633
512	404.2265524	-959.6406314
512	404.2264697	-959.6406304
512	404.2372232	-959.6406293
512	404.2373273	-959.6406281
512	404.2373474	-959.6406278
512	404.2374459	-959.6406265
512	404.225957	-959.6406238
512	404.2258676	-959.6406226
512	404.2258516	-959.6406224
512	404.2380733	-959.6406181
512	404.2254011	-959.6406161
512	404.225225	-959.6406135
512	404.2252114	-959.6406133
512	404.2250885	-959.6406114
512	404.238589	-959.6406104
512	404.2386036	-959.6406102
512	404.2386841	-959.6406089
512	404.2387435	-959.640608
512	404.2247894	-959.6406068
512	404.2388312	-959.6406066
512	404.2388315	-959.6406066
512	404.2388518	-959.6406063
512	404.2389071	-959.6406054
512	404.2389165	-959.6406052
512	404.224605	-959.6406038
512	404.2390359	-959.6406033
512	404.1853793	-959.6405942
466.5256879	422.1146724	-959.6405897
441.2278485	457.3770859	-959.6405828
512	404.1946667	-959.6405737
443.0022874	460.1495304	-959.6404555

Pythonを操るFCMPプロシジャ
～SASとPythonの融合～

○関根　暁史
（藤本製薬株式会社）

The FCMP Procedure making SAS and Python talk to each other.

Satoshi Sekine
Fujimoto Pharmaceutical Corp.

要旨

SAS9.4M6 後期版より proc FCMP の中に Python コードを直接書けるようになり、それを実行できるようになった。本論文では SAS の中での Python の具体的な操作法を紹介する。

キーワード： PC-SAS，proc FCMP，Python

1. はじめに

SAS9.4M6(2019 年 5 月リリース版)より PC-SAS とローカルの Python(Anaconda も可)が直接接続できるようになった。具体的には proc FCMP の中で Python を実行することとなるが、SAS から Python に引数を渡したり、Python の戻り値を SAS が回収して利用できたりするので、SAS と Python がシームレスに連携できるようになっている。インターフェースが SAS だけで完結できる点も魅力である。また SAS ログの中に Python のログも表示されるので比較的デバッグはしやすい。これにより SAS の分析結果を Python がグラフ化・帳票化したり、その逆も可能となった。本論文では、UUID・チェックサムの獲得、フォルダ構成の取得、Excel での ROC 曲線の描画、機械学習と予測値の算出、PDF の Word への変換、ワードクラウド作成、RPA を用いての Word フォントの調整を紹介する。

但し SAS と Python を接続させるためには以下の Windows 環境変数の初期設定だけは必要となる。

```
x 'setx MAS_M2PATH "C:¥Program
Files¥SASHome¥SASFoundation¥9.4¥tkmas¥sasmisc¥mas2py.py"';

x 'setx MAS_PYPATH
"C:¥Users¥[user_name]¥AppData¥Local¥Programs¥Python¥Python38¥python.exe"';
```

上記 2 文を X ステートメントにて 1 回実行頂くのみで、あとは半永久的に Python に繋がることとなる。

2. proc FCMP の使い方

1 章の初期設定さえ済ませてしまえば、proc　FCMP の中で簡単に Python を動かすことができる。proc　FCMP の基本構文の中に、半角スペース 4 個分のインデントは必要ではあるが、Python コード(赤枠)を直接書き込むことができる。以下は UUID(汎用一意識別子)を発生させているプログラムである。UUID とは ISO 規格であり Ver.1～Ver.5 までの方法があるが、ここでは名前(例:あいうえ)から発生させることが可能な Ver.5 を採用している。

```
proc fcmp;
length MyResult $100;
declare object py(python);
submit into py;
def PyProduct():
    """Output: MyKey"""
    import uuid
    id=str(uuid.uuid5(uuid.NAMESPACE_DNS,"あいうえ"))
    return id
endsubmit;
rc=py.publish();
rc=py.call('PyProduct');
MyResult=py.results['MyKey'];
put MyResult;
run;
```

上記プログラムを回転させれば SAS アウトプットに必ず「9607a189-fe1a-5c88-8b73-de3f3d17f6ec」が得られることとなる。

3. チェックサム値の獲得

SAS にチェックサムを獲得する関数は存在していないが、Python にはデフォルトで装備されている。以下は、SHA256 アルゴリズムによってファイルのチェックサム値を獲得している例である(巻末のプログラム 1 参照)。

```
"""Output:"""
import hashlib
with open(infile,'rb') as f:
    id=hashlib.sha256(f.read()).hexdigest()
```

SAS のマクロ変数を Python コードの中に書くことはできないが、SAS のマクロ変数を引数として Python に渡すことはできる。引数(ここではファイルパス)の渡し方であるが、以下のフロー図が模式的に表している。青線が SAS のマクロ変数の流れを示しており、赤線が Python での引数の流れを示している。日本語の引数を渡すためには unicodec 関数などで事前に UTF-8 に変換させておく必要はある。

```
%let _infile=./submissionunit.xml;

proc fcmp;
・・・
def PyProduct(infile,outfile):
・・・
      with open(infile,'rb') as f:
            checksum=hashlib.sha256(f.read()).hexdigest()
・・・
rc=py.publish();
rc=py.call('PyProduct',"&infile.","&outfile.");
run;
```

一方、Python の戻り値の回収の方法であるが、proc fcmp を関数化してしまえば、SAS データセットとして回収することができる。赤線が Python のリターン値の流れを、青線が SAS データの流れを示している。日本語の戻り値は文字化けしてしまうので、全角文字を使いたいときは Unicode 版 SAS を使用することが望ましい。

```
proc fcmp outlib=work.fcmp.pyfuncs;
function MyFunc(arg0 $) $200;
・・・
      """Output: MyKey"""
      with open(infile,'rb') as f:
            id=hashlib.sha256(f.read()).hexdigest()
      return id
・・・
Result=py.results['MyKey'];
return(Result);
endfunc;
run;
```

FCMP の関数化

```
options cmplib=work.fcmp;
data _null_;
   checksum=MyFunc("&infile.");
   put checksum;
run;
```

関数の呼び出し

上記の方法でファイルのチェックサム値を獲得することができた。チェックサムとはファイルが持つ以下のような固有の符号のことである。

「9da675731dcedd0ebada74254bccdff4a8090b8caa36c1329d749558593cd1d2」

4. フォルダ構成の取得

フォルダ構成が作成者の意図したものになっているか、また正しい作成日のファイルが格納されているか確認することとする。SAS ならば X ステートメントや PIPE エンジンなどを使用してフォルダ構成を取得することとなると思うが、Python の glob 関数を用い、オプションを recursive=True とすれば非常に楽にサブフォルダまでたどることができる(詳細は Web 掲載のプログラム参照)。

```
def PyProduct(inpath,outpath):
    """Output:"""
    import openpyxl,os,datetime,glob,saspy,pandas as pd
    from natsort import natsorted
    files=natsorted(glob.glob(inpath+"/**",recursive=True))
```

以下は、eCTD のフォルダ構成を SAS データセットとして獲得した例である。

m1	2023/06/14				
		jp	2023/06/14		
				m1-01-01.pdf	2023/02/22
				m1-02-01.pdf	2023/02/22
				m1-02-02.pdf	2023/02/22
				m1-03-01.pdf	2023/02/22
				m1-03-02.pdf	2023/02/22

右横の数字はフォルダ・ファイル作成日を示しているが、これをファイルサイズや読み取り専用属性情報に簡単に切り替えることは可能である。また SAS データセットの形で取得しているので、これを元に Excel や RTF 出力にするのも容易である。

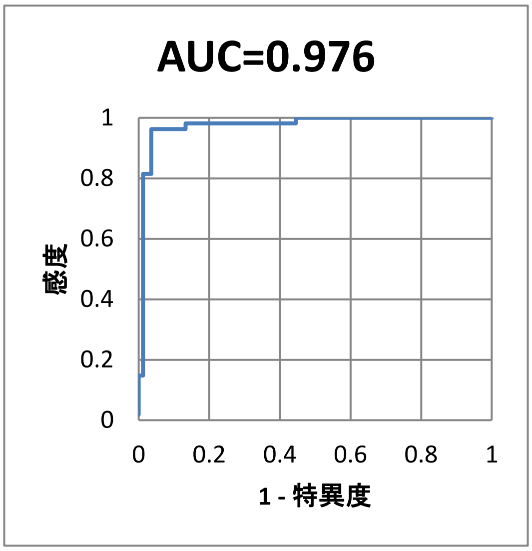
5. Excel で ROC 曲線を描く

SASにもExcelグラフを描く機能(MSCHART)が一時期実装されていたこともあったが、今では廃止されてしまっている。PythonのOpenPyXLを用いれば、Excelの機能を使わずにExcelグラフを描くことができる。下図は、SASのLogisticプロシジャの分析結果をPythonに渡して描画したROC曲線である(詳細はWeb掲載のプログラム参照)。ROC曲線情報は永久SASデータセットで渡し、AUC(曲線下面積)情報は引数の形で渡して記入した。

```
proc logistic data=sashelp.bmt;
  model status=t/outroc=out.rocdat;
run;
  ・・・
df=SAS7BDAT(outpath+"/rocdat.sas7bdat",encoding="sjis").to_data_frame()
df.to_excel(outpath+"/"+outfile)
wb=load_workbook(outpath+"/"+outfile)
chart=ScatterChart()
chart.title="AUC="+auc
```

SAS データセット

引数



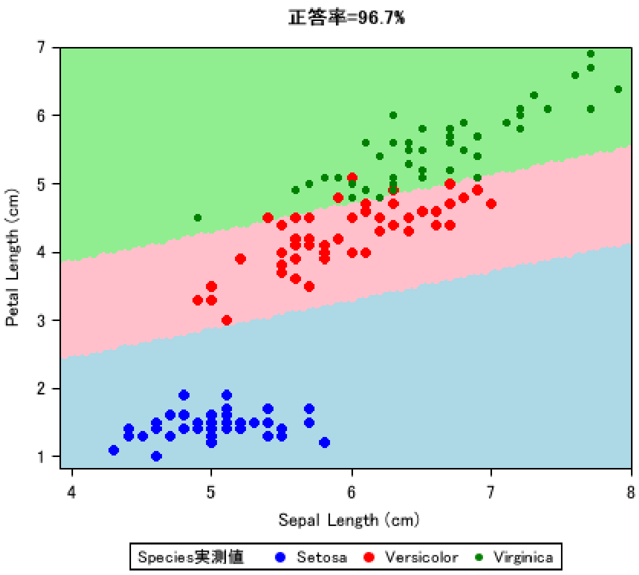
上記は、分析を SAS で行ってグラフ化は Python で行った例である。

6. 機械学習と予測値の算出

SAS Viya がなければ機械学習をすることは困難であろう。しかし Python の力を借りれば機械学習は可能となる(詳細は Web 掲載のプログラム参照)。

6-1. ニューラルネットワーク分析

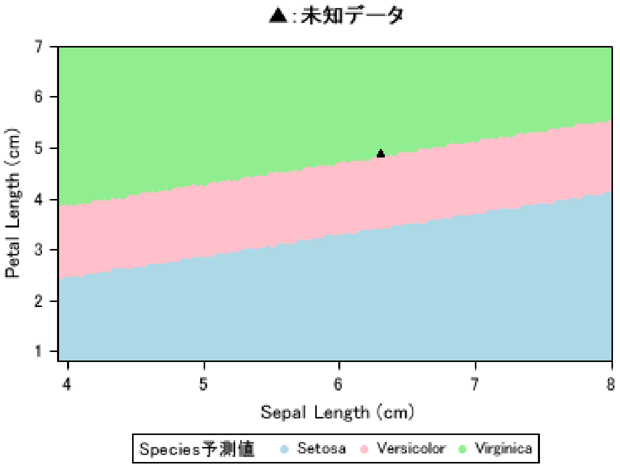
scikit-learn を用い、フィッシャーのアヤメデータのうち SepalLength, PetalLength の 2 変数を説明変数、Species を目的変数、隠れ層 3 層(10,2,10)としたニューラルネットワーク分析を行った。図の背面に描かれているのが予測地図である。この予測地図と色が異なる実測値のドットが誤答である。誤答が 5 つあったので、このニューラルネットの正答率＝145/150×100＝96.7%であった。本解析の予測式(neural_model.sav)は別途保存しておく。



```
from sklearn.neural_network import MLPClassifier
from sklearn.datasets import load_iris
iris=load_iris()
data_train=iris.data[:, [0,2]]
target_train=iris.target
clf=MLPClassifier(hidden_layer_sizes=(10,2,10),random_state=0,max_iter=10000)
clf.fit(data_train,target_train)
filename=outpath +"/neural_model.sav"
```

6-2. ニューラルネットの予測式を未知データに当てはめる

(SepalLength, PetalLength)＝(6.3cm, 4.9cm)で Species が判らない未知データがあったとする。これを **6-1** 章の予測式(neural_model.sav) に当てはめたとき、Species は Virginica 種と予測された。もちろん予測地図上では緑色の箇所にいる。




```
import numpy as np,pandas as pd,saspy,pickle
from sklearn.neural_network import MLPClassifier
x=pd.read_sas(outpath +"/miti.sas7bdat")
data_test=x.values
filename=outpath+"/neural_model.sav"
clf=pickle.load(open(filename,'rb'))
pred_test=clf.predict(data_test) #モデルの当てはめ
```

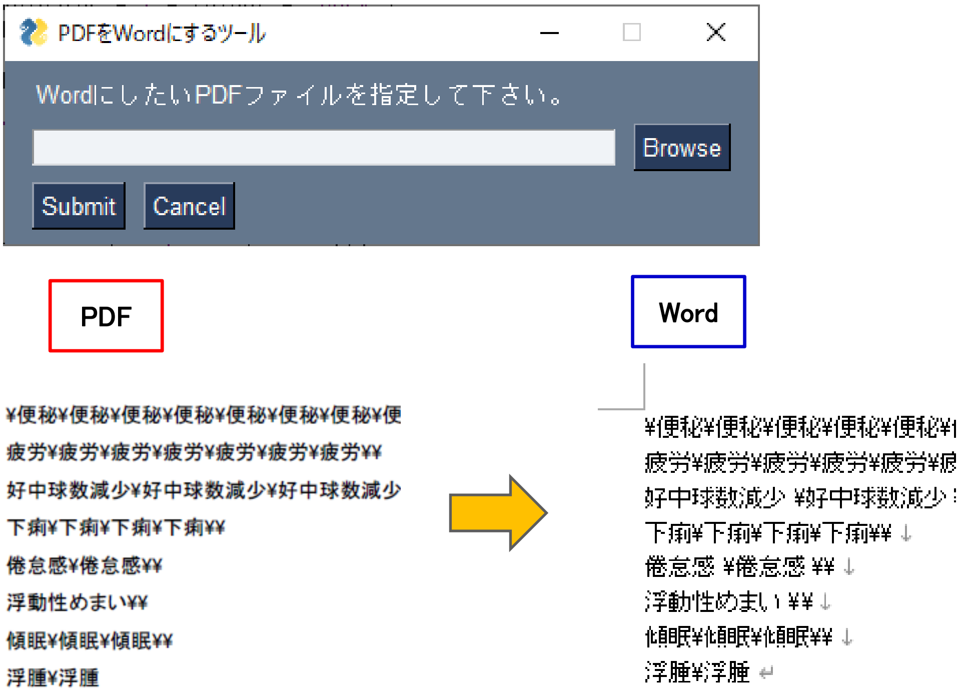
上記は、SASのダミーデータセットをPythonに渡し、Pythonで分析を行い、グラフ化はSASで行うというパターンであった。

7. PDF から Word への変換

Word ファイルから PDF ファイルに変換することはあっても、PDF ファイルから Word ファイルに変換するのは珍しい。PDF ファイルを Word ファイルに変換することによって、SAS 等でより活用し易い形態となるだろう。

7-1. PDF ファイルから Word ファイルへの変換

PyPDF2 を使えば、PDF ファイルの情報を Word ファイルに書き写すことができる(詳細は Web 掲載のプログラム参照)。「変換」というよりも「書き写す」という表現が相応しいと思われるが、Word ファイルにセクション区切りを入れる仕様としているので、PDF のページ番号と Word のページ番号とは一致させることができる。ツールはファイル選択をし易いよう GUI 化している。




```
elif event=='Submit':
    inpath=values['file1']
    paths=os.path.dirname(inpath)
    folder=os.path.split(inpath)[1]
    reader=PyPDF2.PdfReader(inpath)
    page_number=len(reader.pages)
    document=Document()
    i=0
    text=""
    for page in reader.pages:
        i=i+1
        document.add_paragraph(page.extract_text())
        if i != page_number:
            document.add_section()
```

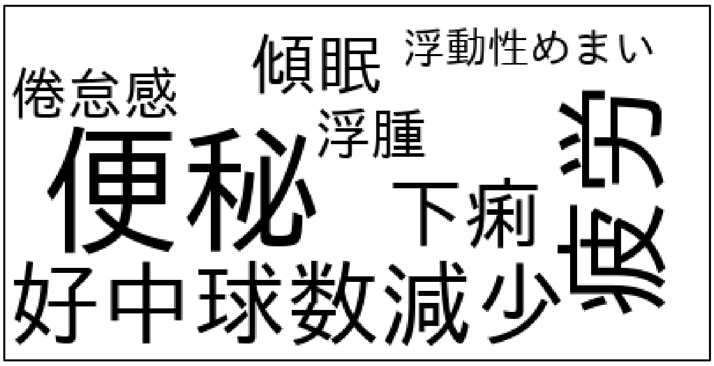
7-2. ワードクラウドの作成

7-1 章のプログラムは、PDF で獲得した情報から同時にワードクラウドを作る機能も有している。ワードクラウドとは、文章やテキストから単語の出現頻度にあわせて文字の大きさを変えて視覚化した

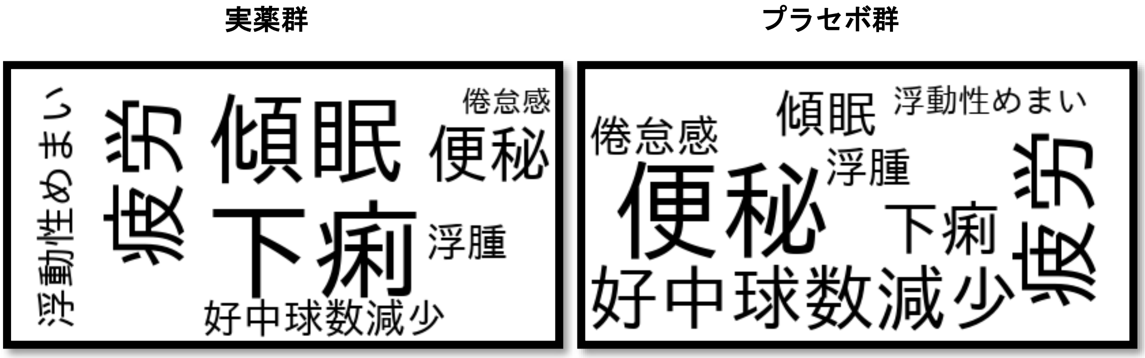
グラフのことである。WordCloud モジュールを使って上記 PDF 情報からワードクラウドを作成した。

```
wordcloud=WordCloud(max_font_size=80,background_color="white",color_func=color_func,font_path=ttf,width=400,height=200).generate(text)
wordcloud.to_file(paths +"/"+ folder +".png")
```

　下図のような有害事象発現頻度に応じて文字が大きくなるワードクラウドが得られた。便秘と疲労が多い傾向にあることが判る。



　上記は発現頻度に応じて文字の大きさを変化させたが、有害事象発現割合に応じて文字の大きさを変えることによって群間比較を表現して見てはいかがだろうか。



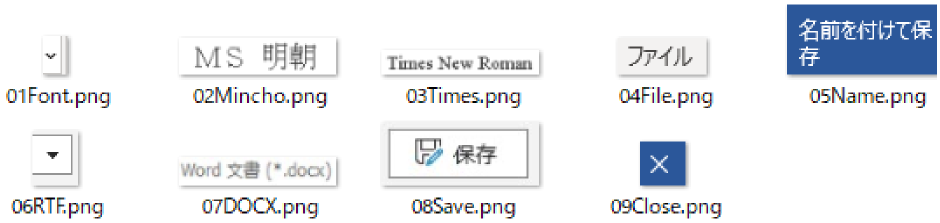
これはダミーデータではあるが、プラセボ群に対する実薬群の副作用として下痢と傾眠が多そうだと一目で判る。

8. RPA による Word のフォント変換

　eCTD の英数字フォントは Times New Roman を使用することとなっている。だが VBA でフォント変換を行ってもギリシャ文字やローマ数字はＭＳ明朝となってしまう。手動操作では Times New Roman に変えることができるので、この手動操作を RPA(Robotic Process Automation)に代替してもらうこととした。PyAutoGUI がボタンを叩く順番は下図のフローの通りである。RPA を行うためにはこのようなアイコンを事前に準備しておく必要がある。

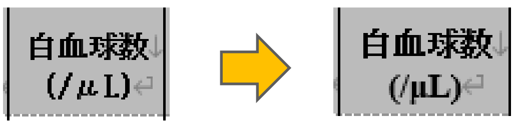
　RTF はファイルサイズが重すぎたり、臨床で直接扱う形式ではなかったりするので、工程の最後に RTF をワードファイルに変換して保存する機能を追加してある(詳細は Web 掲載のプログラム参照)。

本プログラムは SAS が出力した REPORT ファイルをそのまま連続的にフォント変換してしまうことを想定している。




```
from time import sleep
import pyautogui as pgui,os
sleep(8) #8 秒待機
os.chdir(var1)
while True:
    c1=pgui.locateCenterOnScreen("01Font.png")
    if c1:
        break
pgui.moveTo(c1)
pgui.click()
pgui.hotkey("Enter")
pgui.hotkey("ctrl","a")
・・・
```

RPA により μ(マイクロ)を Times New Roman 表示とすることができた。



9. まとめ

SAS9.4M6 後期版以降、SAS と Python がシームレスに対話できるようになった。proc FCMP 自体を直接 SAS マクロ化したり、SAS マクロ変数を Python コードの中に仕込んだりすることはできないが、引数や戻り値の形でやりとりできるようになっている。 これにより、PC-SAS が余り得意としていない機械学習やワードクラウド作成といった機能を Python に補填してもらうことが可能となった。また一元的に SAS インターフェースの中で操作できたり、SAS ログの中に Python のログも表示されたりするのは好まれるであろう。

本論文では SAS と Python を連携させることにより、UUID・チェックサムの獲得、フォルダ構成の取得、Excel での ROC 曲線の描画、機械学習と予測値の算出、PDF の Word への変換、ワードクラウド作成、RPA を用いての Word のフォント調整をすることを紹介した。proc FCMP の構文は初見ではやや

難解ではあるが、1 つプログラムを成功させてしまえば慣れるであろう。
　その他、PDF ファイルの結合・抽出・回転、Excel ファイルの保護と解除、SAS の実行ログファイルのエラーチェックなど多彩な用途に利用できそうだ。

参考文献

1) Michael Whitcher (2019). What’s New in FCMP for SAS 9.4 and SAS Viya，SAS Institute Inc. SAS3480-2019

2) Isaiah Lankham and Matthew T. Slaughter (2023). Friends are better with Everything: A User's Guide to PROC FCMP Python Objects in Base SAS，PharmaSUG 2023 - Paper AP-049

3) SAS Help Center: PROC FCMP Python オブジェクトの使用

付録

```
/*「プログラム1」*/

*---対象ファイルの置き場所指定(パス記入);
%let _inpath=.;
*---対象ファイル名(拡張子ごと記入);
%let _infile=submissionunit.xml;
*---チェックサムファイル出力フォルダー指定(パス記入);
%let _outpath=.;
*---チェックサムファイル名(拡張子ごと記入);
%let _outfile=sha256.txt;

filename _inpath "&_inpath.";
filename _outpath "&_outpath.";
data _null_;
  inpath=tranwrd(pathname("_inpath"),"¥","/")||"/&_infile.";
  call symputx("infile",unicodec(inpath,'utf8'));
  outpath=tranwrd(pathname("_outpath"),"¥","/")||"/&_outfile.";
  call symputx("outfile",unicodec(outpath,'utf8'));
run;

proc fcmp;
declare object py(python);
submit into py;
def PyProduct(infile,outfile):
    """Output:"""
    import hashlib
    with open(infile,'rb') as f:
        checksum=hashlib.sha256(f.read()).hexdigest()
    ff=open(outfile,'w',encoding='utf-8')
    ff.write(checksum)
    ff.close()
endsubmit;
rc=py.publish();
rc=py.call('PyProduct',"&infile.","&outfile.");
run;
```

Base SASによる2次元の半空間深度の実装

田中　真史

（イーピーエス株式会社）

Implementation of the Two-Dimensional Halfspace Location Depth in Base SAS

Masashi Tanaka

EPS Corporation

要旨

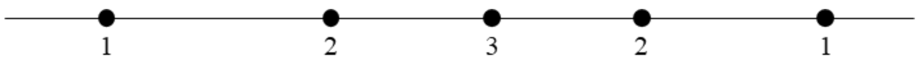
2 次元の測定値について、半空間深度（Halfspace Location Depth）を計算して、散布図に等高線を引いた。この図は、測定値の全体的な特徴を把握することに役立ち、同様のアイデアを発展させた図が、箱ひげ図を 2 次元以上へ一般化したバッグプロットである。本稿では、Base SAS のみを用いて半空間深度を計算し、その性質と課題を考察した。

キーワード：半空間深度（Halfspace Location Depth）、Base SAS、散布図、バッグプロット

1. 緒言

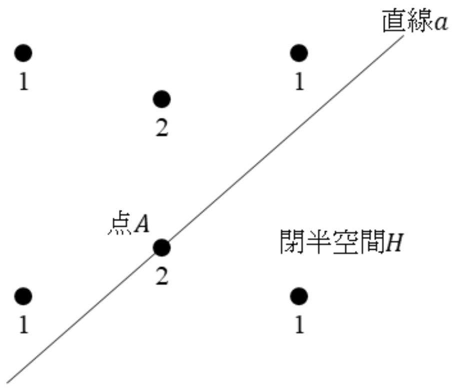
測定値の中央値やパーセント点を算出する手続きを具体的に考えると、数直線上に配置した測定値を外側から順番に眺めていって、条件に合う点を選び出していることに気が付く。この操作は、1 次元の測定値の「深度」を調べていると解釈できる。図 1.1 に示した 5 個の測定値では、一番外側の測定値は深度 1、ひとつ内側の測定値は深度 2、もうひとつ内側の測定値は深度 3 と順に深くなる。

図 1.1　1 次元の測定値の深度



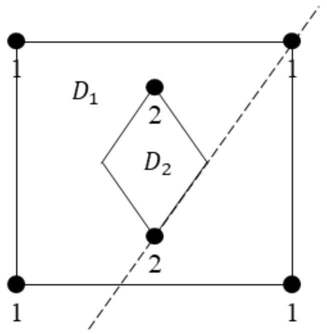
深度は、2 次元（以上）の空間に対して容易に拡張でき、これを半空間深度（Halfspace Location Depth 又は Tukey Depth）と呼ぶ。定義は 2 章に回すが、図 1.2 では、点Aを通る直線aで定まる閉半平面Hに含まれる測定値の数がふたつだから、点Aの半空間深度は 2 である。半空間深度は、測定値以外の（数学的な意味での）点に対しても定義できることに注意する。

図 1.2　2 次元の測定値の半空間深度



半空間深度が 1 以上の点の集合を深度 1 のバッグ（D_1）、2 以上の点の集合を深度 2 のバッグ（D_2）と呼ぶ。図示すると、深度 1 のバッグが深度 2 のバッグを包む入れ子構造の絵が描ける（図 1.3）。

図 1.3　深度 1 のバッグと深度 2 のバッグ



1 次元の中央値やパーセント点の算出を深度の算出と読み替え、もう少し精密に 1 次元の箱ひげ図を 2 次元以上に拡張した図は、バッグプロットと呼ばれている[1]。図 1.3 はバッグプロットではないが、すでに 2 次元の散布図の全体的な特徴を示している。本稿では、これを Base SAS のみで作図する。SAS のバージョンは、9.4M7（日本語版）を用いた。

2. 半空間深度とバッグの定義

2 次元平面上の点xの半空間深度$ldepth(x)$とは、xを含む閉半平面を色々と選んだときの、閉半平面に含まれる測定値の数の最小値である。すなわち、測定値の集合をZとして、xを含む閉半平面をHとすると半空間深度は、

$$ldepth(x) \coloneqq \min_{H} \#(Z \cap H)$$

である。深度dのバッグD_dを、

$$D_d = \{x | ldepth(x) \geq d\}$$

と定義する。一番外側の測定値の深度は 1 で、深度 1 以上の点からなる領域（深度 1 のバッグD_1）の境界上の点（測定値とは限らない）の深度も（ほとんどの場合）1 である（正確には付録を参照）。深度 1 のバッグ

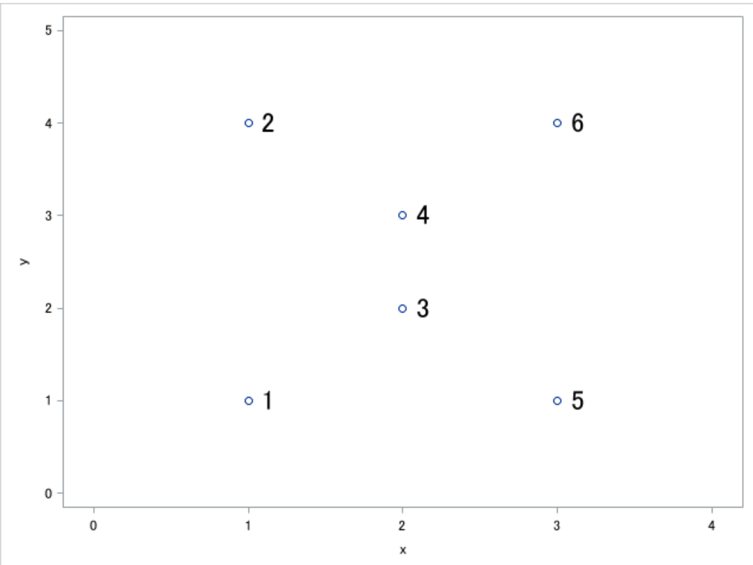
はすべての測定値を含み、その形状は測定値を頂点とする凸多角形になる。一般に、バッグの境界は凸多角形で、測定値はバッグの頂点である。ただし、逆が成立せず、深度 2 以上のバッグについては、前章の図 1.3 のように、凸多角形のすべての頂点が測定値とは限らない。深度が大きくなるにつれてバッグは小さくなり、最終的に、測定値の「中心」付近に到達することができる。この操作は、1 次元の測定値から中央値を探索する操作に類似しており、これが「Depth Median」のアイデアである[1]。

3. 半空間深度の実装

6 個の測定値からなるデータセット XY（コード 3.1）を例にして、半空間深度を実装する。コード 3.1 の下に、データセット XY の散布図を示した。測定値の横の数値は ID 番号である。本実装では、3 個以上の測定値が一直線上に存在する場合を考慮していない。この難点については、結語で触れる。

コード 3.1　実装例に用いるデータセット

```
data xy;
   input id x y;
   cards;
1 1 1
2 1 4
3 2 2
4 2 3
5 3 1
6 3 4
;
run;
proc sgplot data = xy noautolegend;
   scatter x = x y = y;
   text x = x y = y text = id / textattrs = (size = 16pt);
   xaxis values = (0 to 4);
   yaxis values = (0 to 5);
run;
```



準備として、FCMP プロシジャを用いて、測定値の ID から X 座標の値及び Y 座標の値を戻り値として返す関数 X(.)及び Y(.)をユーザー定義した（コード 3.2）。本実装では、FCMP プロシジャと HASH オブジェクトを組み合わせることで、データステップの見通しを良くした。FCMP プロシジャの FUNCTION ステートメントから ENDSUB ステートメントの部分はマクロ化した。

コード 3.2　測定値の座標を取得する関数

```
%macro f(name, dat, var);
   function &name.(id);
      declare hash h(dataset: "&dat.", duplicate: 'e');
      rc = h.definekey('id');
      rc = h.definedata("&var.");
      rc = h.definedone();
      if h.find() = 0 then rt = &var.;
      return(rt);
   endsub;
%mend f;
options cmplib = _null_;
proc fcmp outlib = work._f.f;
   %f(x, xy, x);
   %f(y, xy, y);
run;
options cmplib = work._f;
```

後の処理のために、異なる測定値からなる 30 個（$= 6 \times 5$）の順序対について、測定値から測定値へ直線を引き、直線とすべての測定値の関係を調べた。直線をID_1の測定値からID_2の測定値に引くものとし、左側の閉半平面に含まれる測定値の個数を数えた（ID_1の測定値は除く）。この個数を、やや紛らわしいが、順序対の「深さ」と呼ぶことにする。ID_2の測定値からID_1の測定値へ逆方向に直線を引く場合は、もう一方の閉半平面を考えることになる。閉半平面に測定値が含まれているかどうかは、2 次元ベクトルの外積（$a \times b \coloneqq a_x b_y - a_y b_x$）の値が0以上かどうかで判定した（コード 3.3）。実装では、コード 3.3 を応用して、指定した深さを与える順序対を選出した（コード 3.4）。結果のデータセットは各コードの下に示した。外積の値を 1E-10 で四捨五入したのは、数値を比較する際に浮動小数点誤差の影響を避けるためである。

コード 3.3　順序対で定まる閉半平面に含まれる測定値の個数（下線部が外積による判定）

```
data xy_path;
   if 0 then set xy nobs = nof;
   do id1 = 1 to nof;
      x1 = x(id1);
      y1 = y(id1);
      do id2 = 1 to nof;
         x2 = x(id2);
         y2 = y(id2);
         depth = 0;
         do i = 1 to nof;
            discr = (x2 - x1)*(y(i) - y1) - (y2 - y1)*(x(i) - x1);
            if round(discr, 1e-10) >= 0 and id1 ^= i then depth + 1;
         end;
         if id1 ^= id2 then output;
      end;
   end;
   stop;
   keep id1 id2 depth;
run;
```

	id1	id2	depth
1	1	2	1
2	1	3	4
3	1	4	2
4	1	5	5
5	1	6	3
6	2	1	5
7	2	3	4
8	2	4	2
9	2	5	3
10	2	6	1
11	3	1	2
12	3	2	2
13	3	4	3
14	3	5	4
15	3	6	4
16	4	1	4
17	4	2	4
18	4	3	3
19	4	5	2
20	4	6	2
21	5	1	1
22	5	2	3
23	5	3	2
24	5	4	4
25	5	6	5
26	6	1	3
27	6	2	5
28	6	3	2
29	6	4	4
30	6	5	1

コード 3.4　指定した深さ（ここでは 2）を与える順序対

```
%macro depth(indat, outdat, d);
   data &outdat.;
      if 0 then set &indat. nobs = nof;
      do id1 = 1 to nof;
         x1 = x(id1);
         y1 = y(id1);
         do id2 = 1 to nof;
            x2 = x(id2);
            y2 = y(id2);
            depth = 0;
            do i = 1 to nof;
               discr = (x2 - x1)*(y(i) - y1) - (y2 - y1)*(x(i) - x1);
               if round(discr, 1e-10) >= 0 and id1 ^= i then depth + 1;
               if depth = &d. + 1 then leave;
            end;
            if depth = &d. then output;
         end;
      end;
      stop;
      keep id1 id2;
   run;
%mend depth;
%depth(xy, xy_path, 2);
```

	id1	id2
1	1	4
2	2	4
3	3	1
4	3	2
5	4	5
6	4	6
7	5	3
8	6	3

説明のために、しばらくは深度 2 のバッグのみに着目する。続いて、深さ 2 を与える順序対で決まる直線の交点（PX 及び PY）を算出した（コード 3.5）。このとき、バッグの外側に発生する不要な交点を外積で判別して取り除いた。

コード 3.5　2 直線の交点の算出（下線部が座標の計算）

```
%macro points(indat, outdat);
   data &outdat.;
      set &indat. nobs = nof;
      do i = _n_ + 1 to nof;
         set &indat.(rename = (id1 = id3 id2 = id4)) point = i;
         x1 = x(id1); y1 = y(id1);
         x2 = x(id2); y2 = y(id2);
         x3 = x(id3); y3 = y(id3);
         x4 = x(id4); y4 = y(id4);
         det1 = (x2 - x1)*(y3 - y4) - (x3 - x4)*(y2 - y1);
         det2 = (x3 - x1)*(y3 - y4) - (x3 - x4)*(y3 - y1);
         if round(det1, 1e-10) ^= 0 then do;
            px = x1 + det2/det1*(x2 - x1);
            py = y1 + det2/det1*(y2 - y1);
            flg = 1;
            do j = 1 to nof;
               set &indat.(rename = (id1 = _id1 id2 = _id2)) point = j;
               _x1 = x(_id1); _y1 = y(_id1);
               _x2 = x(_id2); _y2 = y(_id2);
               discr = (_x2 - _x1)*(py - _y1) - (_y2 - _y1)*(px - _x1);
               if round(discr, 1e-10) > 0 then flg = 0;
               if flg = 0 then leave;
            end;
            if flg = 1 then do;
               px = round(px, 1e-10);
               py = round(py, 1e-10);
                  output;
            end;
         end;
      end;
      keep px py;
   run;
   proc sort data = &outdat. nodupkey;
      by px py;
   run;
   data &outdat.;
      set &outdat.;
      id + 1;
   run;
%mend points;
%points(xy_path, pt);
```

	px	py	id
1	1.75	2.5	1
2	2	2	2
3	2	3	3
4	2.25	2.5	4

ここで、コード 3.2 を用いて、関数 X(.)及び Y(.)を PX 及び PY を戻り値とする関数として再定義する。今度は、交点のデータセット PT について、コード 3.4 のマクロを「%depth(pt, pt_path, 1)」と用いて、深度 1 からなる外周を与える点の組を取得した（コード 3.6）。

コード 3.6　2 直線の交点の深さ 1 の順序対

```
options cmplib = _null_;
proc fcmp outlib = work._f.f;
   %f(x, pt, px);
   %f(y, pt, py);
run;
options cmplib = work._f;
%depth(pt, pt_path, 1);
```

	id1	id2
1	1	3
2	2	1
3	3	4
4	4	2

上記のデータセットは、深度 2 のバッグを描画する際に、点を 1 → 3 → 4 → 2 → 1 とつなぐことを示している。この様な処理は、HASH オブジェクトで容易に実装できる（コード 3.7）。

コード 3.7　描画用のデータセットの準備

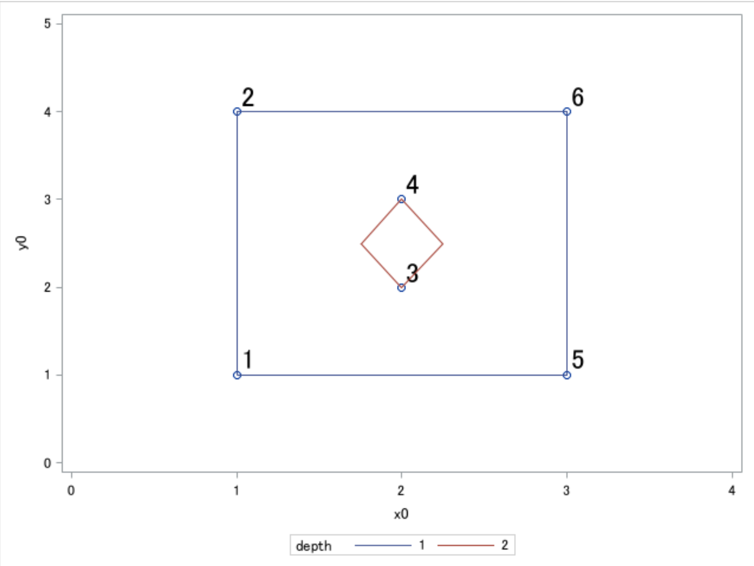
```
%macro mgcl_banana(indat, outdat, depth);
   data &outdat.;
      if _n_ = 1 then do;
         dcl hash h(dataset: "&indat.", duplicate: 'e');
         h.definekey('id1');
         h.definedata('id2');
         h.definedone();
      end;
      set &indat.(obs = 1);
      depth = &depth.;
      id_chain = id1;
      x = x(id_chain);
      y = y(id_chain);
         output;
      do while(1);
         h.find(key: id_chain);
         id_chain = id2;
         x = x(id_chain);
         y = y(id_chain);
            output;
         if id_chain = id1 then leave;
      end;
      keep depth id_chain x y;
   run;
%mend mgcl_banana;
%mgcl_banana(pt_path, chain_d2, 2);
```

	depth	id_chain	x	y
1	2	1	1.75	2.5
2	2	3	2	3
3	2	4	2.25	2.5
4	2	2	2	2
5	2	1	1.75	2.5

実行用にまとめたのがコード 3.8 のマクロで、深度 1 及び 2 のバッグについて境界を算出した。実行結果で、前章の図 1.3 が再現できた。測定値の数が 100 個で、深度 1、2 及び 5 のバッグを描画する例も掲載した（コード 3.9）。

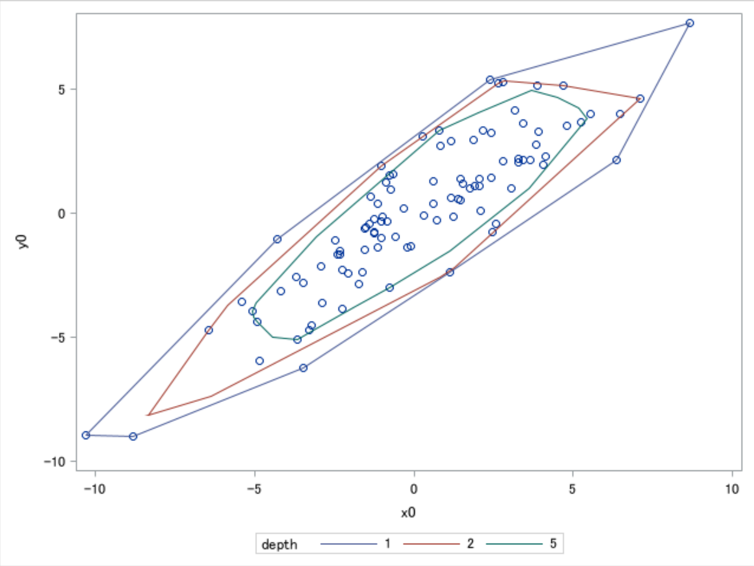
コード 3.8　実行用のマクロ

```
%macro execute(depth);
   options cmplib = _null_;
   proc fcmp outlib = work._f.f;
      %f(x, xy, x);
      %f(y, xy, y);
   run;
   options cmplib = work._f;
   %depth(xy, xy_path, &depth.);
   %points(xy_path, pt);
   options cmplib = _null_;
   proc fcmp outlib = work._f.f;
      %f(x, pt, px);
      %f(y, pt, py);
   run;
   options cmplib = work._f;
   %depth(pt, pt_path, 1); *depth = 1;
   %mgcl_banana(pt_path, chain_d&depth., &depth.);
%mend execute;
%execute(1);
%execute(2);
data out;
   set xy(rename = (x = x0 y= y0)) chain_d:;
run;
proc sgplot data = out;
   scatter x = x0 y = y0 / datalabel = id datalabelattrs = (size = 16pt);
   polygon x = x y = y id = depth / group = depth;
   xaxis values = (0 to 4);
   yaxis values = (0 to 5);
run;
```



コード 3.9　100 個の測定値からなるテストデータ

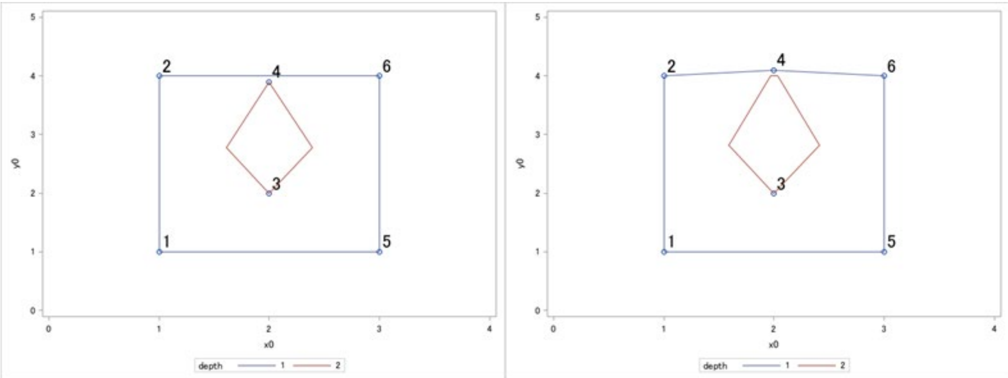
```
data xy;
   seed = 15678;
   do id = 1 to 100;
      z1 = rannor(seed);
      z2 = rannor(seed);
      z3 = rannor(seed);
      x  = 3*z1 + z2;
      y  = 3*z1 + z3;
         output;
   end;
   keep id x y;
run;
```



4. 結語

半空間深度を Base SAS で実装する際の難点は、データセット中の複数のオブザベーションを組み合わせた処理が頻発することである。そのため、FCMP プロシジャや HASH オブジェクトといった強力な手法が必要になった。ループ処理を多数回行うため、計算量が多くなる問題があるが、R の mrfDepth パッケージ[2]でも、近似を用いて計算するオプションがあり、半空間深度の計算には、本質的に多くのステップが必要なようである[3]。本稿の実装では、3 個以上の測定値が一直線上に存在する場合を考慮しなかった。この状況は、1 次元でタイデータが存在する状況に類似している。プログラム的には、測定値の座標の値をランダムにわずかにずらすことでエラーを回避することは可能である。しかし、図 4.1 のように、測定値とバッグの関係が定性的に変わる状況であるため、これも根本的に難しい問題であると思う。

図 4.1　3 個の測定値がほぼ一直線上に存在する場合



半空間深度は、SAS/IML による実装例が公開されており、金融分野でストレステストを行う際に、ロバストな解析手法に応用ができるらしい[4]。JMP 15 ではバッグプロットが作図可能で、Web 上で綺麗な図を眺めることができる[5][6]。将来、SAS でも簡単に半空間深度の計算とバッグプロットの作図ができるようになれば、筆者は喜んでそれを活用したい！

付録　バッグの性質

2次元の測定値の有限集合をZ、点xを含む閉半平面をHとして、

$$l(x,H) := \#(Z \cap H)$$

と定義する。xの半空間深度（単に深度と呼ぶ）は、

$$ldepth(x) := \min_{H} l(x,H)$$

である。深度d（≥ 1）のバッグD_dを、

$$D_d = \{x | ldepth(x) \geq d\}$$

と定義する。深度dのバッグについて、以下の性質1–6が成り立つ。性質1は定義から明らかである。性質2、3より、バッグは閉凸集合である。性質4、5は、バッグの境界についての性質である。

1. 深度$d+1$のバッグは深度dのバッグの部分集合である。
2. バッグは閉集合である。
3. バッグは凸集合である。
4. 測定値はバッグの境界上の点である。
5. バッグの境界上の点は、バッグに含まれるとは限らないふたつの測定値を結ぶ線分上に存在する。

3個以上の測定値を通る直線がひとつも存在しないと仮定すると、性質6が得られる。性質6が、本稿の実装においては重要である。

6. 深度dのバッグの境界上の点xの深度はdである。さらに、バッグに含まれるとは限らないふたつの測定値とそれらの測定値で定まる閉半平面Hがとれて、$ldepth(x) = l(x,H) - 1$である。

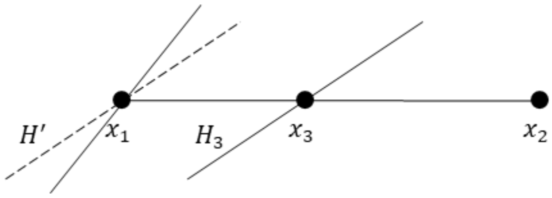
［2の証明］　深度dのバッグの補集合が開集合であることを示す。補集合に含まれる任意の点xを選び、その深度（$< d$）を与える閉半平面をHとする。Hをわずかに平行移動してH'をとることを考える。平行移動の方向はふたつあるが、どちらの方向についても、H'に含まれる測定値の個数が真に増加しないような動かし方が存在する。このとき、H'に含まれる点の深度はd未満である。よって、補集合の任意の点は内点である。すなわち、バッグは閉集合である。

■

［3の証明］　深度dのバッグが凸集合であるとは、そのバッグに含まれる2点x_1、x_2を結んだ線分上の任意の点x_3で、$ldepth(x_3) \geq d$であることを意味する。$H_3 \ni x_1$の場合は、x_3の深度を与える閉半平面H_3をx_1まで平行移動して、下図の破線を境界とする閉半平面H'をとる。

$$ldepth(x_3) = l(x_3,H_3) \geq l(x_1,H') \geq ldepth(x_1)$$

であり、x_1の深度がd以上なので、$ldepth(x_3) \geq d$である。$H_3 \ni x_2$の場合も同様である。



■

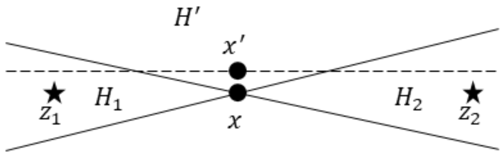
［4 の証明］　測定値zの深度を与える閉半平面をHとする。Hを、Hの内部の方向へわずかに平行移動してH'をとる。H'の境界上にあるzの近傍の点x'では、

$$ldepth(z) = l(z, H) > l(x', H') \geq ldepth(x')$$

のように、深度が真に減少する。これは、測定値がバッグの境界上の点であることを意味する。

■

［5 の証明］　深度dのバッグの境界上の点xでは$ldepth(x) \geq d$である。xの近傍にあるバッグの外点をx'とする。x'の深度を与える閉半平面をH'とすると、$l(x', H') = ldepth(x') < d$である。$H'$は、着目しているバッグと交わらないことに注意する（交わったら、交点で深度がd未満になってしまう）。xに対して下図のように閉半平面H_1、H_2をとると、$l(x, H_1) \geq ldepth(x) \geq d$だから、$H_1 \setminus H'$（差集合）にひとつ以上の測定値が存在することが分かる。それをz_1とおく。同様に、$H_2 \setminus H'$にもひとつ以上の測定値が存在するのでz_2とおく。ここで、点x'を点xに近づけ、H_1、H_2の境界（実線）を、H'の境界（破線）に近づける。すると、xは測定値z_1とz_2を結ぶ線分上に存在することが分かる。



■

［6 の証明］　前半は性質 5 の証明の図を用いる。深度dのバッグの境界上の点xは、測定値z_1、z_2を結ぶ線分上に存在する（性質 5）。仮定より、3 個以上の測定値を通る直線が存在しないので、z_1方向に線分z_1z_2を延長してもz_1以外の測定値は存在しない。よって、

$$ldepth(x) \leq l(x, H_1) = ldepth(x') + 1 < d + 1$$

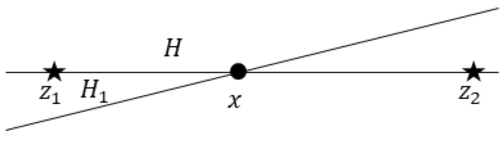
である。バッグの定義より$ldepth(x) \geq d$だから、

$$ldepth(x) = d$$

が得られ、境界上の深度はdである。同時に$ldepth(x) = l(x, H_1)$も得られ、下図のように直線z_1z_2で定まる閉半平面のひとつをHとすると$l(x, H) = l(x, H_1) + 1$なので、

$$ldepth(x) = l(x, H) - 1$$

である。



■

参考文献

1. Rousseeuw, P. J.; Ruts I.; Tukey J. W. (1999). "The Bagplot: A Bivariate Boxplot". *The American Statistician*. 53 (4): 382–387.
   https://www.researchgate.net/publication/247788661_The_Bagplot_A_Bivariate_Boxplot. 2023 年 8 月 31 日閲覧.
2. Package ‘mrfDepth’.
   https://cran.r-project.org/web/packages/mrfDepth/mrfDepth.pdf. 2023 年 8 月 31 日閲覧.
3. Dyckerhoff, R.; Mozharovskyi, P. (2016). "Exact computation of the halfspace depth". *Computational Statistics & Data Analysis*. 98: 19–30.
   https://arxiv.org/pdf/1411.6927.pdf. 2023 年 8 月 31 日閲覧.
4. Donin D.; Korol D.; Tkachenko M. (2011). Scenario and Stress Testing using SAS/IML®.
   http://support.sas.com/resources/papers/proceedings11/353-2011.pdf. 2023 年 8 月 31 日閲覧.
5. JMP® 15 の新機能 最新バージョンの新しい機能と拡張点. SAS Institute Japan 株式会社 JMP ジャパン事業部.
   https://www.jmp.com/content/dam/jmp/documents/jp/software/jmp/jmp-15/new-features-jmp15-ja.pdf. 2023 年 8 月 31 日閲覧.
6. 2D Box plot (Bag Plot) of Fisher's Iris Data.
   https://public.jmp.com/packages/sqr97V4_nQCS9xxxFjGMr. 2023 年 8 月 31 日閲覧.

世界初のLSD（線形分離可能なデータ）の判別理論

〇新村秀一[1]

（[1]成蹊大学　名誉教授）

The First Discriminant Theory of Linearly Separable Data (LSD)

Shuichi Shinmura

Professor Emeritus of Seikei University

要旨

1971 年から種々の判別データを SAS や JMP で実証研究し 2015 年に **Theory1（新しい判別理論）** を完成した。Springer1[11]は誤分類のある学生と多重共線性のある児頭骨盤不均衡（CPD）データ、そして Iris、複数の試験の合否判定、スイス紙幣の真札と偽札、小型車と普通車の 4 種の LSD を分析した。1995 年に人間の遺伝子の出す**蛋白量(発現量)**を測定した **Microarray**（n 症例<<p 変数: n*p）から「**多変量の癌遺伝子**」を見つける医学研究が行われ、データが公開された。そこで Theory1 の応用としてとり組み、第 1 世代の 6 データが LSD で n 個以下の遺伝子の組の **SM**(Small Matryoshka)や最小次元の LSD である **BGS**（Basic Gene Set）に分割できる事が分かった。さらに LSD の 4 つの普遍的なデータ構造（**Fact3**）を見つけた。そして医学研究者が医学診断に役立つ 2 段階スクリーニング法(**Method3**)を開発し、遺伝子数が 5 個以下で検証標本の平均誤分類確率（ER）の **M2** が 0 である「多変量の癌遺伝子の候補」である **vital BGS** を見つける事に成功した。これらの中に既に医学研究で見つけた 1 変量の遺産的な発癌遺伝子が含まれておれば、vital BGS の医学的な特徴が分かる。これで 2022 年に **Theory2（癌の遺伝子データ解析）** が完成した。これらの成果をまとめて 3 冊目の LSD の判別理論を出版することにした。そこで Theory1 で既に分析したデータを Theory2 の成果で見直すと、全く期待もしなかった LSD の判別理論の新しい発見があった。一番重要なのは、2 例の誤判別症例のある CPD データで、これを省いて LSD を作り判別すれば、多重共線性のない M2 の小さな BGS が得られる。すなわち BGS を求めることが「Occame の剃刀」の要求を満たす最小次元の Best なモデルで、他の変数選択手法がいらない。それ以上にデータに一意に決まる誤分類例を省くと全ての判別データが LSD になり、画期的に良い結果が得られる。これは**変数選択法**より大きな成果をもたらす**ケース選択法**である。この変数選択とケース選択は、判別理論に取って福音となり、今後多くの成功事例が期待できる。

キーワード：169 Microarrays、4 種の通常の LSD、判別データの 4 つの普遍的なデータ構造（Fact3）、LSD の Matryoshka 構造、SM 分割、BGS 分割、DF 分割、世界初の癌遺伝子データ解析理論、横長データ(n<p)

1.　はじめに

筆者は 2015 年に、1991 年以来行ってきた判別理論のデータ解析の研究をまとめて、新しい判別理論を完成した[11]。10 月 28 日に統計シンポジーウムでその成果を発表し、第 1 世代の 6Microarray が公開されていることを知った。Theory1 の応用研究として、**RIP**(最適線形判別関数)で Shipp[15]のデータ(77*7129)を判

別した。すると 1 秒で最小誤分類数(minimum NM,MNM)が 0 で LSD である。32 個の係数だけが非零で、残り 7097 個が 0 になる。これが **SM** と呼ぶ「**癌と正常を分ける多変量の癌遺伝子の候補**」の発見である。そして僅か 54 日で 6 データを判別し LSD でその中に多くの異なった SM を含むことが分かった(LSD の **Structure2**)。

11 月 10 日に JMP ユーザー会[8]で、SVD を使った高次元の LDF と PCA の報告があった。新版を借りて 6 データを判別したところ、ER は 2%から 17%で LSD でないことが分かった。SAS や JMP のような豊富な人材を抱えた統計ソフトの結果は、個人の研究より信頼に値する。多くの統計研究者や数学者が SVD で同じ研究をしているが、JMP が駄目であれば、彼らの研究は成功しないことは明白である。

そこで 2016 年から本格的に「**高次元遺伝子データ解析（Theory2）**」の研究をした。最初の SM1 を省いて残りの遺伝子を判別すると 2 番目の SM2 が求まる。そこで数理計画法の LINGO[9]で全ての LSD である Type-1 の SM を求める Program3 を作った。また Theory1 で銀行紙幣データ(200*6)[4]の 2 変量の(x4, x6)が BGS であることは分かっていたので、Microarray を多くの BGS に分割する **Program4** を作った。

iPS 研究で山中教授は 3 万個以上の人の遺伝子から遺伝子 DB で 24 個の万能細胞ができる遺伝子を特定した。この 24 個から恐らく**逐次変数減少法**と同じ方法で、山中 4 因子を見つけた。すなわち 24 個の遺伝子が SM に対応し、4 因子が BGS と考えれば理解しやすい。しかし Theory2 研究のバズワードとして、3 つの困難が有名だ:1）遺伝子データ解析は高次元のため解析手法がない。2）また適切な変数の組を求めるのは NP ハードである。3）そして多くの雑音の中から信号を見つけるのは難しい、と言っている。

これら全て間違いであることが私の研究が示している。すなわち Microarray は LSD でありその中に SM から最小次元の BGS までの Matryoshka 人形のデータ構造（LSD の **Structure1**）を持っている。Program3 の SM 分割は **Structure2** であり、BGS 分割は **Structure3** である。最尤推定法を使ったロジスティック回帰や重回帰分析は横長データで DF=n と一般に n 個の変数を選ぶ事が **Structure4** である。

統計ソフトを使って実証研究している **Data Scientist** は、**横長データ**（n<p）は、重回帰や LDF や PCA が p 変数から n 変数以下の変数を選ぶ（DF=n）ことを知っている。また連立方程式を解く場合、p 変数から n 変数を選ばないと解が求まらないことは、高校数学の常識である。この事実を理工学研究者がまったく気づいていないということが大きな問題である。すなわち Microarray の判別は、P 変数から N 変数を選ぶ「**組み合わせ理論**」が適している。以上の LSD の汎用的なデータ構造（**Fact3**）が Theory2 の結論である。しかし癌研究者が、癌の遺伝子診断に直ぐに使えない。そこで一般的な「**症例設計の三原則**」に代えて「**癌症例設計の三原則**」を見つけた。そして 100 例以下の大腸癌から多くの **vital BGS**（M2=0 で遺伝子数が 5 個以下）を見つけた。医師がこれらの中に、医学研究ですでに見つけている**遺産遺伝子**が含まれていれば、**多変量の癌遺伝子**という新しい研究成果がえられる［10-13］。2023 年に Theory2 の研究成果を 4 種の普通の LSD と 2 種の誤分類例のあるデータで再分析して、画期的な発見をした。3 冊目の本の概略を以下で紹介する。

2. Theory1（新しい判別理論）と Theory2（高次元遺伝子データ解析）

2.1 Fisher の仮説の問題

Fisher と同世代の統計学者は、頼りない記述統計量を数学的な厳密な理論にしたかった。そこで数学者の Gauss の 2 地点間の繰り返し測定値が Gauss 分布になることと最小二乗法の成果を取り入れて、推測統計学を考えた。大学に職がなかった彼は、ロダムステッド農場の栽培植物の農事研究者になった。そこで実験計画法や分散分析法を考えた。これらの手法は、推測統計学の代表的な手法である。例えば重回帰分析の予測値は、観測値の正規分布の平均になる。これは Gauss 分布を見つけた状況とよく合致している。Fisher が開

発した線形判別関数（LDF）は、医学診断などの多くの応用分野に用いられ役に立った。しかし２群が平均だけ異なる同じ正規分布という仮説を考えれば、単に正規分布の比の対数を取ると簡単に LDF になることで導かれた。またFisherは判別係数や判別結果を表すERの標準誤差は定義していないので推測統計手法でない。この意味で他の理論に比べて見劣りがする。多くの統計手法に最小二乗法が使われる。しかし筆者の研究ではLSDが判別できるのは筆者の開発したRIP（最適LDF）とH-SVM[15]とロジスティック回帰だけである。最尤推定法を用いて 2 群の分布にあうロジスティック回帰が求まり、LSD が判別できると考える。他の全ての判別関数やML（機械工学）の研究者がそれらを拡張したClassfiersで研究しているが、それらの手法は全くLSD には役に立たない。すなわち判別分析は正規分布と最小自乗法でなく、組み合わせ理論と最尤推定法が適している。さらにいえばMicroarrayが高次元の2つの正規分布と考える事自体、現実を無視した考えられない前提である。判別理論と重回帰や分散分析などの正規分布が適した理論とは、区別して理解すべきだ。

2.2　Theory1:判別分析の4つの問題と2つのFactと2つのMethod（1971年～2015年）

統計ソフトを使った長年の実証研究で、判別分析には4つの重大な問題があることが分かった[11]。**問題1**は判別結果を表すERが全く信頼性がないことである。異なる判別手法で同じデータを判別しても、また判別境界を変えても ER が異なる。LSD が判別できるのは、上記の 3 つの LDF だけである。それ以外の LDF はER=0であればLSDであるが、全てのLSDを正しく0にできない問題がある。**問題2**は筆者以外にLSDの研究がない。銀行紙幣データで（x4、x6）がBGSである。BGSを含む16モデルがLSDの信号で、残りの47モデルはLSDではない雑音で分析対象にならない。この後、2個の自明なLSDを見つけた。**試験の合否判定**を2個の得点で 50 点で合否判定する場合、自明な LDF=T1+T2-49.5 で合否判定できる。BGS から、出題者が学生と設問の関係が分かるし、入試の分析も簡単に行える。また普通車と小型車（44＊6）は2個の産業規格の排出量と座席数がBGSで、車種が決まる。この63モデルのM2を、**Program2の10重CV**で直接評価し、MNMとM2の値からSMやBGSが簡単に見つかる。多くの産業製品は、規格が1変数のBGSになるデータがあり、自明なLSDになると考えられる。また Iris データ（150＊4）の 4 変数が 2 組の 1 変量の BGS と 2 変量の BGS である。13個のモデルがLSDになり、2組の1変量のX1とX2がLSDでない。恐らく多くの動植物の種別も、自明なLSDになると思われる。我々のまわりにはLSDが多い。しかしMicroarrayがLSDでなるのは、正常細胞が癌化するとエピジェネチック変化が起きて、癌が正常と完全に分かれるためと考える。すなわち「発現量は人類が出会った最高品質の計測値」である。**問題3**は大学入試センター試験の数学ⅡBで、ERが30%を超えるものが見つかった。また合格者がある設問を全員正解し、不合格者がばらつく場合にQDFが合格者全員を不合格に誤判別する。これらは分散共分散行列の重大な瑕疵である。**問題4**は判別分析にはERや判別係数のSEがない。そこで**10重CV（Theory1）**を開発した。一般にはLahenbruch[7]が貢献したLOOかそれを拡張したK重CVが用いられるが、これらは統計の母集団と標本の関係を無視している。**Theory1**はJMPで元のデータを10回コピーし検証標本（擬似的な母集団）と考える。そして乱数で並べ替えて10分割して学習標本とする（検証標本からのサンプル）。これで10組のRIPを計算し、検証標本を10回判別する。そして検証標本の10組の平均のERを**M2**と呼ぶ。このERの分布から判別係数のSEが求まるが、このM2の値が最小のモデルを一番良いと評価するのが**Validation1**である。そしてTheory2で一番役に立った。

この他、**Fact1** で「誤分類数NMと判別係数の関係」を判別係数の空間で説明した。また「MNMの単調減少性(MNM_k>=$MNM_{(k+1)}$)」の **Fact2** をみつけた。以上の詳しい説明は、昨年度迄の発表を見てほしい。重要なのはProgram1でRIPやH-SVMとソフトマージンSVMという4種のLDFを選ぶ事ができる。H-SVMはLSDでしか計算できないので、Penalty c=1000ぐらいのS-SVMでLSDを判別できることもある。そしてRIPでShippのデ

ータを判別し SM1 が求まった。SM1 を省いた残りの遺伝子を判別し SM2 を手作業で行い SM10 迄求めても終了しない。そこで **SM 分割する Program3** を開発した。

2.3　Theory2：第 1 世代の 6 データと第 2 世代の 73 データ（2016 年～2022 年）

2.3.1　第 1 世代の 6 Microarray（2016 年～2018 年）

Theory1 で完成していた **Program1**（RIP）と **Program2**（10 重 CV）に加えて、2016 年には **Program3**（SM 分割）と **Program4**（BGS 分割）を完成した。そして 6 データで **Fact3** を確認した。しかし Ward 法と PCA で SM を分析しても 2 群が綺麗に分かれない。そこで、分割した k 組の Type-1 の SM から、k 個の RIP 等の判別スコアを p 個の遺伝子の代わりに変数とした**信号データ**(n*k)を作った。重要な事は、各 Microarray に対し 1 組の統計分析しやすい小標本である。それらを統計分析した結果を **Springer2**[12]で紹介した。

2.3.2　第 2 世代の 73 データ（2019 年～2022 年）

2019 年に国際的な科学技術の SNS の **Research Gate**(https:// www.researchgate.net /profile /Shuichi-Shinmura)から、バイオ工学の Buruno ら[2]が 2007 年以降の第 2 世代の Microarray を登録した CuMiDa という DB を開発したメールが届いた。彼らは、5 種の 1 群と、正常のある 57 の 2 群と、3 から 7 群の 16 の Microarray を提供している。彼らはデータの品質試験を行い、問題のある症例を修正あるいは削除している。統計では行っていない、一種のケース選択である。78 データを 8 種の Classifier で、3 重 CV で分析して ER の優劣を論じている。そして多くのデータで kernel-SVM とランダム・フォレスト（RF）が良いとしている。一方イタリアの Cilia[3]らは、種々の変数選択法で変数の組を選んで、10 重 CV で DT の ER が 8%以上で良いという成果を報告している。彼らの研究は多くの理工学研究の中では秀逸である。

筆者は、ER を LSD の信頼性から 5 分類した。**分類 1** は 3 種の LDF は、LSD の ER の信頼性が一番高い。それ以外の LDF は ER＝0 の場合は LSD であるが、ER に信頼が置けず**分類 2** である。kernel－SVM や QDF 等の非線形判別関数は**分類 3** である。多くの ML 研究者が提案する MLP,Naïve Bays，RF は ER=0 であっても LSD と断定できないので**分類 4** である。**分類 5** は心電図診断で用いられた DT（枝別れ）や介護保険システムの構築に用いられた分類木（JMP の Partitioning）が**分類 5** である。2 群判別は分類 1、多群の階層ある実システムの構築には分類 5 を用いるべきというのが私の結論である。東大医学部の伝説の三秀才と言われた開原名誉教授が厚生省で「介護保険システム」を企画された。東京都の病院の副院長であった D 医師が、介護病院などで集めた「1 分間タイム Study」と呼ばれる Big データを SPSS の訳の分からない難しい手法で分析し、開発に遅れていた。私が分類木（JMP の Partitioning）を勧めた。彼は 3 か月ほどで CHAID で分析したアルゴリズムを C に置き換えて全国システムを完成した。

Theory2 の理工学研究の問題は、比較している手法に LDF が無く根拠無く kernel-SVM が良いと誤解し、分析手法の対象を無節操にクラスター分析まで広げている。そして意味の無い個々のデータ毎に ER で優劣を論じている。また変数選択や FS も AI などの画像診断で成功した間違った役に立たない手法である。

筆者の研究では、ER の質が一定で科学技術の基本である LSD であるか否かを基準にしている。その結果、LSD の判別には RIP と H-SVM とロジスティック回帰しか使っていけないことを示した。さらに、MNM=0 で LSD である上に、10 重 CV で SM や BGS の M2=0 が多変量の癌遺伝子の候補とした点である。この LSD である事実を指摘した研究が一つも無い事と、横長データの分析の特殊性を知らず、多くの研究者が間違った研究を行ってきた。科学技術の研究史上の最大の信じられない不祥事である。

筆者は 1 群の 5 データは、教師無しのクラスター分析しかできず研究対象外である。何の目的でこれらのデータを集め研究しているのか理解できない。さすがに正常のある 57 の 2 群は、医学研究の基本である 1)

正常と癌を比較研究する（原則 1）、2）症例数はほぼ同数が望ましい（原則 2）、3）症例数は 100 以上と多いほど良い（原則 3）、という「**症例設計の 3 原則**」が重要だ。しかし検査費用が高いので、100 例以下の原則 3 を満たさないデータが多い。3 群から 7 群は、正常のないデータも多い。しかし目をつぶり、これらから 2 群の組み合わせで 106 の 2 群データを作った。計 163 のデータで 6 種の古いデータで見つけた Fact3 を確認した。驚くことに Program2 で SM と BGS を検証し、幾つかの SM と BGS が M2=0 になり、Ward 法と PCA の散布図で 2 群に分かれた(**Validation3**)。「全ての遺伝子が正しい組み合わせで Type-1 の SM と BGS になる。理工学研究者のいう「多くの雑音から一組の信号を見つけることが困難」という言い訳が間違いである。

筆者は遺伝子に関する初歩的な知識の勉強をしたが、生物学や医学の沢山の事実関係が頭の中に定着しない。そこで次に自分流の研究の考え方をまとめた。

1. **データ解析は学際的な学問**である。計測値が遺伝子の出す蛋白量である限り、DNA であろうが RNA であろうが正しく分析できる。現時点では、Golub らの研究に NIH が研究資金を出さなかったことによる Golub ショックで、Microarray はだめで RNA-seq という新しい計測値に癌研究がシフトしている。しかし Microarray で失敗した研究方法で、青い鳥を探して RNA-seq で分析しても成功しない。既に 169 種の Microarray で Fact3 という成果が出ている。RNA-seq の品質が良いのなら、学際的なデータ解析技術でもっと良い結果が得られるだけだ。
2. 正しい統計と数理計画法の知識とソフトの利用で、実データを徹底的に調べた実証研究が重要である。多くの失敗した研究者は、真の Data Sientist で無かったことが原因である。
3. 小手先の医学知識でなく、医学常識を優先すべきである。例えば癌は遺伝子の病気であり、不均質な病気である。たった一組の 50 個程度の多い遺伝子の組を選択して癌診断ができると考える Golub ら[5]、Shipp ら、Singh ら、Alon ら[1]、Tien らの第 1 世代の医学研究は間違いである。
4. 高次元の遺伝子から多変量の発癌遺伝子を選択する研究は、理工学研究の使命である。そして選んだ遺伝子の組を厳しく真の多変量の癌遺伝子と決定できる根拠のある評価法が必要である。このため 4 種の Validation を考えた。医学研究者が重視する Genome Cohort と生存時間解析（**Validation5**）はこれらを行った最後に行うべきである。
5. 医学的な知識は必要ないが、医学的な成果である 400 個以上の 1 変数の遺産的な発癌遺伝子が既にある。得られた vital BGS がこれらを含んでいるかを検討する **Validation4** で、未知の「多変量の癌遺伝子の候補」の多変量の癌遺伝子を特徴づけられる。含んでいない vital BGS は、医学研究が見つけていない新しい多変量の癌遺伝子の候補か否かは、分からない。
6. 誰もが使える分析方法を最終的には 2 段階スクリーニング法にまとめた。個人の研究者が PC 環境で研究できるように、分析の処理単位として遺伝子を 12,761 個の組に分割した。そして 1 組の分析が約 2 日で結論が出せるようにした。他の研究者が平行して行えば、研究者は 1 ヶ月以内に簡単に多変量の癌遺伝子の組で研究を行える。現在、遺産遺伝子の組を医学的に考察し用いているので、そこそこの成果が出る。しかし選んだ遺伝子をロジスティック回帰で判別すれば、LSD でないことは直ぐに分かる。

2.3.3　第 2 世代の 2 群以上の 73Microarray の評価（2021 年）

169 の Microarray による Fact3 は、データ解析による成果である。そこで、医学研究に真に役立つために Method3(2 Step Screening Method)を開発した。そして「**症例設計の 3 原則**」の原則 3 がおかしく**新原則 3**（100 例以下が良い）ことを 3 研究で確かめた。

Study1 は、次の 3 データを分析した。**Liver3**(181/176)は 181 人の癌と 176 人の正常者が含まれ患者設計

の 3 原則を満たすが、図 1 の左図に示す散布図に多くの誤分類例がある。M2 の範囲は[4.03%, 15%]である。右図の Breast6(101/15)は原則 2 を満たさないので散布図(Validation3)は悪い結果を示し、M2 も悪い。

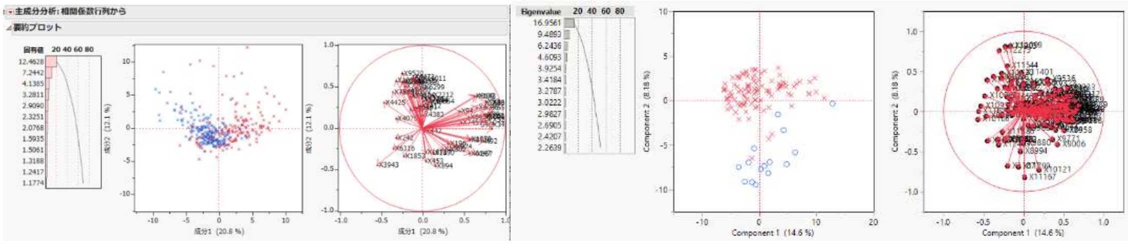


図 1　左は Liver3、右は Breast6 の散布図

Colorectal6(31/32)は、新原則 3 を満たす。図 2 は SM1 の分析結果を示すが、左の Ward 法と右の散布図は綺麗に 2 群が分かれる。M2 の範囲は[0%, 21.27%]で、多くの M2=0 の BGS がある。1 検体の計測にお金がかかるので、Breast6 のように正常例の検査数を少なくしたり、正常例のない 3 群以上のデータも多い。癌の遺伝子診断だけが他の医学診断と研究方法と目的が異なっているようだ。

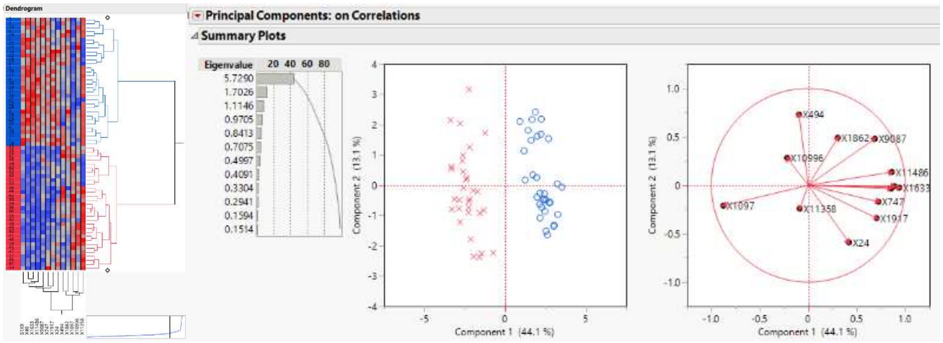


図 2　Colorectal6 の SM1 の Ward 法（左図）と PCA の散布図（右図）

そこで Study2 で Colorectal6 に条件の似た Colorectal5(26/26)[6]を分析した。Validation1 と Validation3 で、図 3 の散布図に示す 4 例の誤分類があり、10 重 CV の M2 の評価も悪い。この 4 例を省いた LSD を分析すると、Colorectal6 よりも多い BGS がある。従って、「**癌患者設計 3 原則**」は、原則 2 にかわって 60 例程度（100 例以下）の中程度の標本サイズが適しているに変更した。

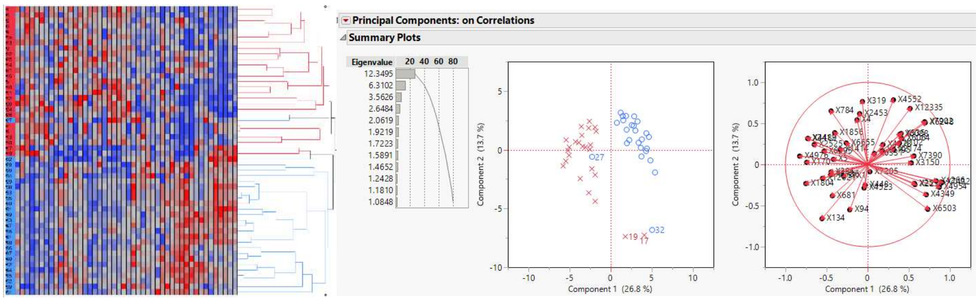


図 3　Colorectal5 の SM1（52 ＊ 46）の Ward（左）と PCA（右）

Study3 では症例数の多い Liver3 から散布図で問題のある症例を省いて 300, 200, 100, 60, 40, 30 症例を分析し、60 例や 40 例が良い結果を示すことを確認した。そして、2 群以上の 73 データを新 3 原則を満たす Priority-1 と、原則 2 を満たさない Priority-2 の 44 組をを検討した。Colorectal5 の 52 例と 48 例を比較すると、誤分類を省いた後者が **1,743 組の vital BGS** があることが分かった。これは多いようにみえるが、医師が **Validation4** で既存の遺産癌遺伝子が含まれる vital BGS を選ぶだけで、従来の医学研究のように 1

変量の遺産遺伝子を医学的な検討で組み合わせて選ぶことより短時間で正しい結果が得られる。

2.4　Theory3：新しい LSD の判別理論（2023 年）

Theory3 では、4 つの通常の LSD と 2 組の誤分類例のある通常のデータを Theory2 の結果で見直し、次のような予想外の素晴らしい結果が得られた。

2.4.1 CPD データ（240*190）

CPD データは、180 例の帝王切開手術を受けた妊婦と 60 例の自然分娩例である。日本医科大学の鈴村教授は、児頭骨盤不均衡の症状を検討して帝王切開手術か自然分娩かの簡単な診断法を提案した。17 の測定値から計算される 2 変数の X9=X7-X8 と X12=X13-X14 で、これら 6 変数間に 2 個の強い**多重共線性**がある。図 4 は逐次変数増加法（FJ と FS）と減少法（BJ と BS）で選んだモデルで、JMP と SAS で求めた LDF の NM を示す。FJ と BJ は JMP、FS と BS は SAS の結果である。両図とも変数増加法で 9 変数前後まで変数が増えると NM は減少するが、それ以降に多重共線性に関係する変数がそろうと増加する。一方減少法は 19 変数から 8 変数から 10 数まで NM は増えて、多重共線性を解消する変数が省かれると減少傾向になる。ただし SAS と JMP は計算方法の違いによる傾向の違いがある。この様に多重共線性のある変数を解消する決定法がないので、1979 年頃に半年以上を費やして解決した。

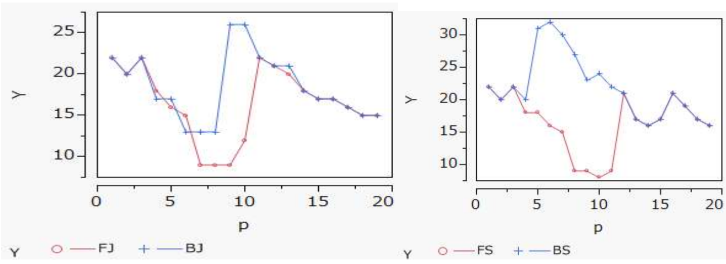


図 4　JMP (左) と SAS (右)の変数増加法（赤）と減少法（青）の NM

Program1 の RIP で、CPD240 に固有の 2 例の誤分類が見つかった。これを省略しテスト標本として扱い、CPD238（LSD）を作成する。Program3 は、SM1(18 変数)を見つけた。Program4 は、BGS1(14 変数)と Type-2 BGS2（5 変数)を見つけた。Program2 で 19 変数、SM1、BGS1、BGS2 などの 4 モデルを 10 重 CV で検証した。M2 はこれらの 4 モデル(BGS1、SM1、19 変数、および BGS2)を評価し(Validation1)、それぞれ 0.002%、1.028%、1.028%、および 12.54%になる。この結果は、変数の少ない BGS1 が M2 の評価で他よりも優れていることを示す(Occamz Razor/Principle of Parsimony)。

通常のデータ(n>=p)を分析する場合、次の単純な 4 ステップである。この分析はわずか半日で終る。

1）RIP はデータを判別し誤分類がある場合は、修正した LSD を作成する。これは**変数選択方法**よりも有用な適切な**ケースの選択法**になる。

2）Program3 は Type-1 の SM とほぼ 1 個の Type-2 の SM を見つける。

3）Program4 は Type-1 の BGS とほぼ 1 個の Type-2 の BGS を検出する。

4） Program2 は、全 BGS を全ての SM と全変数と評価する。現時点では、BGS1 が最小の M2 値になる。すなわち BGS1 は、Occamz Razor を満足する最もコンパクトで最高の判別結果のモデルになる。

5）もし銀行紙幣データ、日本車データのように 6 変数であれば、Program2 で 63 個のモデルの M2 を直接評価して、M2 と MNM の値から SM や BGS を決めることができて非常に簡単である。

RIP、3 つの Fact、および 4 種の LINGO プログラムは、全ての人に新たな判別分析の新世界を開く。一度これまで分析したデータがあれば、試しに分析してみれば、直ぐに納得できるだろう。

2.4.2 学生のアンケートデータ(40*6)

誰もが変数の意味を理解できるため、このデータを SAS、Statistica、SPSS、JMP を使用して日本語の初級の統計教科書を 4 冊書いた。このデータは単純ですが、判別超平面の候補である学生 8 人が月額 50,000 円を費やしているため、8 つの判別結果が異なる。それらは、RIP、改訂 LP-OLDF、S-SVM (ペナルティ c=10000)、Fisher の LDF、ロジスティック回帰、QDF、RDA と DT です。RIP で 5 人の誤分類された生徒を省いて LSD を作成し、35 人の生徒を分析した結果、単純かつ明確な結果が得られた。

2.4.3 スイス紙幣データ (200*6)

本データは、2 変量の BGS1 の(X4、X6)がある。BGS を含む 16 モデルは LSD(Signal)で、M2 の範囲は [0、0.4%]です。他の 47 モデルは LSD ではなく、M2 の 47 モデルは 0.5%を超えている。BGS1 の M2 は 0% で、最もコンパクトな 2 変数で、M2=0 の最良のモデルです。Theory2 では、Microarray で Validation2 の正しさと有用性を証明できなかった。

2.4.4 試験の合否判定データ(約 130*100)

合否判定を T1 と T2 の 2 個の得点で合格点が 50 点の場合、LDF は自明の f(T)=T1+T2-49.5 になり、自明の LSD になる。合格者は f(T)>0 で判断でき、不合格者は f(T)<0 で判定でき、全ての試験は自明の LDF で LSD になる。また、合否判定は Microarray と同じデータ構造である。

多くの試験結果のうち得点分布の 90%を合格点とする結果を示す。100 の質問(M2=10.5%)を、難易度でもって T1(29 変数、M2=9.1%)、T2(12 変数、M2=9.1%)、T3(19 変数、M2=6%)、T4(40 変数、M2=8.7%)に 4 分類した。Program3 は、Type-1 SM1(77 変数、M2=9.2%)と Type-2 の SM2(23 変数、M2=11%) を検出した。Program4 は、Type-1 BGS1 (8 変数、M2=0%)、BGS2(11 変数、M2=2.7%)、BGS3(13 変数、M2=4.4%)、BGS4(16 変数、M2=4.1%)と BGS5(15 変数、M2=4.3%)と Type-2 BGS6(37 変数、M2= 12%)を検出した。この結果は、重要な真実を示す。

1) BGS1 は最もコンパクトなモデル(8 変数)で、Occamz Razor を満たす最良のモデル(M2=0%)です。

2) 100 問は悪い判別モデルです。恐らく、ほとんどの教師は、全ての質問で生徒の理解を測ることができると誤解している。

3) 私の 4 分類も悪いモデルです。この分類に基づいて試験を設計し、学生には T1、T2、T3、T4 の順に高いレベルの理解が必要であると考えていたが間違っていた。

4) ただし T3 は SM より良い結果になり、2 個の SM は合否判定には役立たない。この結果は、私が Theory2 で SM の分析に多くの時間を費やしたことが間違いであることを示す。山中先生のように、私も BGS の研究にもっと時間を割くべきでした。ただし、10 個の SM から BGS を取得すると計算時間が短縮されるため、Method3 では SM が役立つ。

5)教師が Program4、Program3、Program2 の使い方を理解していれば、テストや入学試験を正しく公平に分析・評価できる。教育工学の強力な武器になる。

2.4.5 Fisher のアヤメのデータ

Irisデータ(150*4)は3種の種別があり、(setosa、virginica)、(setosa、vercicolor)、および(setosa、virginica & vercicolor)の 3 つの判別データを分析した。散布図からこれらが LSD であることを容易に分るため、Springer1 では LSD として統計分析しなかった。しかし x3 と x4 は 2 個の一変量の BGS で、(x1,x2)は 1 個の二変量の BGS で、2 個の一変数 x1 と x2 だけが LSD でない。他の 13 モデルは 3 つの BGS の組み合わせの LSD です。また、3 つの判別データ毎に 14 の Matryoshka 人形になる。これで、Matryoshka

/Nested データ構造 (Structure1)の意味が詳しく理解できる。恐らく全ての動物と植物の種が同じ一変量の BGS の測定値を持つ自明な LSD と考える。すなわち身の周りから容易に LSD を発見できる。

2.4.6　第 4 章の日本自動車データ(44*6)

日本では、排出量 x1 と座席数 x3 の 2 個の工業規格で、普通車と小型車が定義されている。x1 と x3 の 2 個の単変量の BGS がある。Program2 で全 63 モデルを評価した。BGS を含む 44 モデルが LSD で、19 モデルが LSD ではない(MNM>0)。1 変数の X3 と X1 を持つ Type-1　BGS1(M2=0)と BGS2(M2=0)、3 変数(X1、X2、X6)の SM2(M2=0)を見つけた。従って変数が少ない場合、Program2 で 2 組の BGS と 1 組の SM1 が分かる。x1 を含む 12 モデルは RIP=5.92*X1-4.893 になる。x3 を含む 20 モデルは、RIP=2*X3-9 になる。他の LDF は異なる係数になる。恐らく工業製品の工業規格が一変量の BGS になると考える。

3　結論：Data Science を支える SAS であってほしい

1970 年頃から遺伝子データの研究をしている Harvard 大学の Golub 教授らの 4 プロジェクトが 2004 年までに第 1 世代の Microarray を研究し、従来の医学研究で顕微鏡と生物学的知見を用いて見つけた発癌遺伝子や癌の抑制遺伝子（**遺産遺伝子**）と異なる方法を摸作した成果を発表し Microarray を公開した。統計や ML や AI などの理工学研究者の中に高価なデータを無料で使えるので高次元遺伝子解析の研究を行った。理工学研究のテーマは、多次元の Microarray の多くのノイズの中から信号である多変量の癌遺伝子の候補を見つけること、そして判別関数や Classifier で判別して k 重 CV 等で ER を比較し、個別データ毎に手法の優劣を示すことにある。これらの研究は筆者の得た結果に比べて、余りにもお粗末で近年の科学技術研究の最大の不祥事である。

筆者の研究で、第 1 世代と第 2 世代の 169 の Microarray が全て LSD である。そしてそれらが SM や BGS や DF という n 個以下の遺伝子の組で LSD に分割できる。僅か数個の遺伝子が MNM が 1 以上の雑音の SM や BGS になる。これらが真の多変量の癌遺伝子の候補であるかを検証する 4 種類の Validation を考えた。その結果、第 1 世代の 6 データは、Validation1 で M2=0 になるデータが無く、Validation3 で Ward 法と PCA で 2 群に分かれなかった。それが第 2 世代の 163 データの中に、M2=0 になる SM や BGS があり、散布図でも 2 群は分かれた。そこでこれらの BGS を多変量の癌遺伝子の候補と呼ぶことにした。この事実から、第 2 世代のデータは品質が向上したと考える。そして LSD を判別できるのは 3 種の LDF しか無い事を示した。これは従来の研究が行っている判別と変数選択（あるいは FS）の手法が役に立たないことを示す。

以上から理工学研究が関与できる研究として、Fact3 を見つける 4 つのプログラムで Theory2 は完成したと宣言できる。しかし見つけた BGS が医学に役立つためには、医学的な検証が必要になる。この時点で理工学研究の範疇を外れるが、誰にでも使える 2 段階スクリーニング法(**Method3**)を考えた。そして医学研究で既に見つけている遺産遺伝子で評価する事にした。すなわち M2=0 で遺伝子数が 5 個程度以下の **vital BGS** を見つける。医学研究者がこの中に遺産遺伝子が含まれておれば、その特徴で多変量の癌遺伝子の特徴を決めれる。含んでいない vital　BGS は、全く新しい発癌遺伝子か、偶然に検証結果を通り抜けたいずれかになる。このため、従来の医学研究の常識である「**症例設計の 3 原則**」を 100 例以下の中規模の症例の方が良いとする「**癌症例設計の 3 原則**」に改めた。これは**発現データが人類がであった最高品質の計測値**である。それは恐らく正常細胞が癌細胞になると起きるエピジェネティク変化のためと考えている。そして Colorectal5 で素晴らしい結果を得た。

多くの理工学研究者が失敗したのは、彼らは先行研究に書かれた数式を追い、実際の目の前にある

Microarray をデータの科学のための統計ソフトと数式で表される対象の科学である数理計画法で実証研究する真の Data Scientist で無かったからである。また高次元データを横長データと理解せず、連立方程式の解の条件や、横長データに対して重回帰や LDF の DF が n になることを知らなかった。

近年癌診断にも AI が活躍している。癌の撲滅には種々のアプローチが望ましい。その点で人間の知識の範囲で、個人研究者が自前の研究環境で、データ解析による癌の遺伝子診断のためのスクリーニング法を開発した。Data Scientist の中から AI に負けない人知でこの最重要テーマに挑戦してほしい。そして SAS や JMP はそれを支援してほしい。癌に限らず、一般の医学診断、あるいは動物の医学診断、さらにデータに群情報を与えて判別データにする事で、これまで得られなかった新分野を開拓できる。

文献

1. Alon U et al. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Cancer and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc. Natl. Acad. Sci. USA, 96: 6745-6750
2. Bruno CF, Eduardo BC, Bruno IG, Marcio D (2019) CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology. 26-0: 1-11
3. Cilia ND et al. (2019) An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets. Information 10, 109: 1-13
4. Flury B, Riedwyl H (1988) Multivariate statistics: a practical approach. Cambridge University Press, New York.
5. Golub TR et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999 Oct 15, 286/5439: 531-537
6. Hinoue T, Weisenberger DJ, Lange CP, Shen H et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 2012 Feb;22(2):271-82. PMID: 21659424
7. Lachenbruch PA, Mickey MR (1968) Estimation of error rates in the discriminant analysis. Technometrics. 10 (1): 11.
8. Sall JP, Creighton L, Lehman A (2004) JMP Start Statistics, Third Edition. SAS Institute Inc.
9. Schrage L (2006) Optimization Modeling with LINGO. LINDO Systems Inc.
10. Shinmura S (2000) A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics, 5: 133-142
11. Shinmura S (2016) New Theory of Discriminant Analysis After R. Fisher. Springer, Dec.
12. Shinmura S (2019) High-dimensional Microarray Data Analysis. Springer, Dec.
13. Shinmura S (2021a) First Theory of Cancer Gene Data Analysis of 169 Microarrays and Four Universal Data Structures for Big Data. CSCI-ISCB: COMPUTATIONAL BIOLOGY1-14. , Springer Nature.
14. Shinmura S (2021b) Twenty-three SeriouSMistakes of Cancer Gene Data Analysis since 1995. CSCI-ISCB: COMPUTATIONAL BIOLOGY1-14. Transactions on Computational Science & Computational Intelligence, Springer Nature.
15. Shipp MA et al. (2002a) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8: 68-74
16. Vapnik V (1999) The Nature of Statistical Learning Theory. Springer.
17. 新村秀一(2004) 『JMP活用　統計学とっておき勉強法』講談社
18. 新村秀一(2007) 『JMPによる統計レポート作成法』丸善
19. 新村秀一(2007) 『最適線形判別関数』日科議連
20. 新村秀一(2007) 『数理計画法による問題解決法』日科議連

SAS ユーザー総会 2023
@東京大学 福武ホール

各種のシグモイド曲線に対する
オフセットを活用した任意のパーセント点の逆推定と 95%信頼区間

高橋 行雄
BioStat 研究所（株）

Inverse estimation of arbitrary percentage points
using offsets for various sigmoid curves and 95% confidence intervals

Yukio Takahashi
BioStat Research Co.,Ltd.

要旨： 各種のシグモイド曲線状の反応が得られたときに，その任意のパーセント点となる *dose* と 95%信頼区間を推定したい．シグモイド曲線の代表例はロジスティック曲線であるが，低用量から反応が急に立ち上がり，その後の反応はゆったりと増加する場合には，ゴンペルツ曲線をあてはめることが望ましい場合がある．その際に，反応の 10%，50%，90%となる *dose* を逆推定したい．ゴンペルツ曲線とは逆に，反応がゆっくり立ち上がる場合には，ワイブル曲線のあてはめが必要となる場合もある．推定される位置パラメータを反応の 10%，50%，90%とするためのオフセット値もそれぞれ異なる．環境ホルモンが子宮に与える影響についての共同研究のデータを用いて，シグモイド曲線の選択法を示し，反応の 10%，50%，90%を推定するためのオフセットの与え方について提示し，推定結果を示す．解析には，SAS の NLIN プロシジャを用いるが，計算原理の理解の助けになるように Excel での計算方法にも言及する．

キーワード：シグモイド曲線，逆推定，オフセット，SAS/NLIN プロシジャ，環境ホルモン，子宮重量

1. はじめに

生物を対象とした実験での反応の多くは，シグモイド曲線状になることが知られている．シグモイド曲線となる関数として累積分布関数が使われてきた．化合物の急性毒性データの 50 パーセント致死量（LD_{50}）を推定するためのプロビット法は，正規分布の累積分布関数が用いられている．正規分布の累積分布関数に代え，ロジスティック分布の累積分布関数も広く使われている．累積分布関数は，0.0 から 1.0 の範囲で単調増加する関数であるが，量的反応に対して，最小反応から最大反応の範囲になるように拡張することができる．

量的反応に対する各種のシグモイド曲線は，応用分野ごとに定式化されてきたので，様々な関数形が用いられており，比較検討する際の障害となる．低い用量で反応が急に立ち上がり，高い用量での反応が緩やかに上昇する場合には，ゴンペルツ曲線をあてはめたい．逆に，低い用量での反応が緩やで，高い用量での反応が急に上昇する場合には，ワイブル曲線をあてはめたい．低い用量でも高い用量でも曲線の形状が同様であれば，ロジスティック曲線のあてはめで済ませたい．得られたデータに対し，どの曲線を用いるべきか統計的に比較検討し，最大反応と最小反応の 10%点，50%点など任意のパーセント点となる用量とその 95%信頼区間を求めたい．

2. 各種のシグモイド曲線のパラメータの共通化

代表的なシグモイド曲線は，正規分布の累積分布関数であり，平均値μと標準偏差σによって規定されていて，0 から 1 へ単調増加なので，最小値をθ_{min}，最大値をθ_{max}となるようにするために

$$y_i = \theta_{min} + (\theta_{max} - \theta_{min}) \cdot F_{NOR}\left(\frac{x_i - \mu}{\sigma}\right) + \varepsilon_i, \quad \varepsilon_i \sim \text{正規分布} \tag{1}$$

のように拡張する．なお，$F_{NOR}()$を標準正規分布の累積分布関数とする．

ロジスティック分布のパラメータは，応用分野ごとに様々であるので，正規分布の場合に合わせたい．ただし，正規分布の標準偏差σは，分散の平方根に等しいが，ロジスティック分布のσ_{LGS}は，分散の平方根ではなく $\sigma^2 = (\pi^2/3) \cdot \sigma_{LGS}^2$ のように，分散の平方根よりも $(\sqrt{3}/\pi = 0.5513)$ と小さいことが知られているので，パラメータとしては，μ_{LGS}とσ_{LGS}を用いてロジスティック曲線を

$$y_i^{LGS} = \theta_{min}^{LGS} + \frac{\theta_{max}^{LGS} - \theta_{min}^{LGS}}{1 + \exp\left[-\dfrac{(x_i - \mu_{LGS})}{\sigma_{LGS}}\right]} + \varepsilon_i \tag{2}$$

のように拡張して定義する．

ゴンペルツ曲線は,

$$1)\ \ y = Kb^{e^{-cx}},\ \ 2)\ \ y = Ka^{b^x},\ \ 3)\ \ y = \alpha \exp[-\beta e^{-kx}],\ \ 4)\ \ y = \exp[-\exp(a - bx)] \tag{3}$$

に示すように様々な式が提示されていが，4) の最大極値分布の式を用い，パラメータを $(a - bx)$ ではなく $(x - \mu_{MEV})/\sigma_{MEV}$ を使うことにする．なお，添え字の MEV は，(*<u>M</u>aximam <u>E</u>xtreme <u>V</u>alue Distribution*）を意味していて，最小値と最大値も区別し，ゴンペルツ曲線を

$$y_i^{MEV} = \theta_{min}^{MEV} + (\theta_{max}^{MEV} - \theta_{min}^{MEV}) \cdot \exp\left[-\exp\left(-\frac{x_i - \mu_{MEV}}{\sigma_{MEV}}\right)\right] + \varepsilon_i \tag{4}$$

のように拡張して定義する．

ワイブル分布は，故障あるいは寿命データの解析で標準的に用いられているのみならず，一般化線形モデルでの 2 値反応に対する解析法として，「補 2 重対数法」のための分布としても知られている．ワイブル分布に基づくシグモイド曲線は，ゴンペルツ曲線とは対照的に，xが小さい場合には反応がゆっくりと上昇し，xが大きくなるにつれて急激に上昇するシグモイド曲線となる．ワイブル分布は，ゴンペルツ曲線の場合と同様に，関数の形式はまちまちであり，日本の信頼性工学分野では,

$$F_{WBL}(t) = 1 - \exp\left\{-\left(\frac{t}{\eta}\right)^m\right\} \tag{5}$$

のようにパラメータにmとηが使われ，mを形状パラメータ，ηを尺度パラメータとしている．尺度 (scale) パラメータηは，実際には「位置パラメータ μ_{WBL}」であり,「形状（shape）パラメータm」の逆数が，σ_{WBL}となる．ワイブル分布の$F_{WBL}(t)$に対し，ηを $\eta = \exp(\mu_{SEV})$，mを$m = 1/\sigma_{SEV}$ とするとことにより，最小極値分布（*SEV* : *<u>S</u>mallest <u>E</u>xtreme <u>V</u>alue Distribution*）の関数形式 $F_{SEV}(\ln(t))$ に変換することができる．さらに，指数と対数をセットで加えることにより,

$$\left.\begin{aligned} F_{SEV}(\ln(t)) &= 1 - \exp\left\{-\exp\left[\ln\left(\frac{t}{\exp(\mu_{SEV})}\right)^{1/\sigma_{SEV}}\right]\right\} \\ &= 1 - \exp\left[-\exp\left(\frac{\ln(t) - \mu_{SEV}}{\sigma_{SEV}}\right)\right] \end{aligned}\right\} \tag{6}$$

が得られ， $x = \ln(t)$ と置き換え，

$$y_i^{SEV} = \theta_{min}^{SEV} + (\theta_{max}^{SEV} - \theta_{min}^{SEV}) \cdot \left\{ 1 - \exp\left[-\exp\left(\frac{x_i - \mu_{SEV}}{\sigma_{SEV}} \right) \right] \right\} + \varepsilon_i \tag{7}$$

式 (2) および式 (3) と同様の形式にすることができる．

3. ロジスティック･ゴンペルツ･ワイブル曲線のあてはめ

山田・吉田・高橋（2003），「陰性および陽性対照があるシグモイド曲線－ダミー変数を持つ非線型回帰モデルの応用―」の事例で，表 1 に示す「環境ホルモン EE（ethinyl estradiol）投与後のラット子宮重量」が取り上げられている．データは，共同研究の施設ごとの平均値が示されている．

表 1　環境ホルモン EE（ethinyl estradiol）投与後のラット子宮重量（blotted uterine）

施設		ethinyl estradiol (EE), μg/kg						
番号	Vehicle	0.01	0.03	0.1	0.3	1	3	10
1	102.35	95	105	112.22	190.45	319.78	373.72	382.00
2	120.82	115	115	123.47	217.48	351.32	384.72	404.32
3	115.92	115	120	144.42	213.95	326.07	378.37	354.37
7	121.62	120	125	131.25	220.83	317.52	387.43	391.67
8	79.22	90	80	105.08	211.13	287.68	262.20	273.73
9	108.47	115	115	123.60	211.37	357.57	353.82	362.05
11	82.45	100	100	113.38	191.23	297.67	307.60	312.40
18	89.25	90	90	91.80	193.07	334.95	334.48	366.20
19	99.17	100	100	83.17	104.67	135.17	234.17	332.67
平均	102.14	104.44	105.56	114.27	194.91	303.08	335.17	353.27
	0.01, 0.03群はグラフから読み取り，0.03, 0.3, 3 群は，0.0316, 0.316, 3.16 の略表示							

Kano, J., Onyon, L., Haseman, J., et al. (2001). The OECD program to validate the rat uterotrophic bioassay to screen compounds for in vivo estrogenic responses: phase 1, Environmental Healthy Perspectives, 109(8);(785-794).Table 5.

この事例に対し，山田ら（2003）は，累積ロジスティック分布関数を拡張したロジスティック曲線が適用されているが, 図 1 に示すように用量反応関係を平均値で概観すると， 0.1μg/kg から 0.3μg/kg に掛けて子宮重量が急激に増加し， 1.0μg/kg 以上では緩やかな子宮重量の増加となっており，ゴンペル曲線（最大極値分布）のあてはめが適切と思われる．Excel のソルバーを用いて 3 種のシグモイド曲線をあてはめ，それぞれの残差平方和の大きさで，あてはまりの性能評価を行なう．

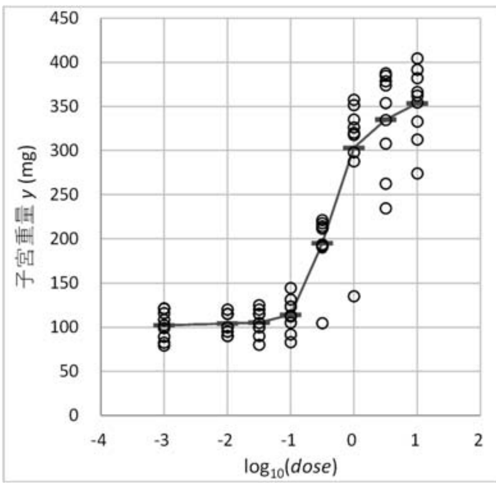


図 1　子宮重量の変化

図 1 から，Vehicle 群の平均値は，低用量群の平均値の延長線上にあると判断されるので，用量段階を 2 段階落とした *dose* $= 0.001\mu$g/kg として解析する. 表 2 左に示すのは, 表 1 の Vehicle を含め 8 用量群×9 施設=72 のデータを行方向に展開し，3 種のシグモイド曲線をあてはめた結果であり，表 2 右には，グラフ表示のために *dose* の間隔を細かく設定し，滑らかなシグモイド曲線が得られるように推定値を計算した結果が示されている．初期値は，図 1 から推定されるパラメータとして共通の $\hat{\theta}_{max} = 350$， $\theta_{min} = 100$， $\hat{\mu} = -0.5$， $\hat{\sigma} = 0.5$ とした．この初期値をそれぞれの曲線のパラメータの欄にコピー&ペーストし，Excel のソルバーで残差平方和 S_e が最小になるようにパラメータを変化させた結果が示されている．

表 2　Excel のソルバーによる子宮重量データに対する 3 種のシグモイド曲線のあてはめ

				MEV	*LGS*	*SEV*					
	初期値	**350**	$\hat{\theta}_{max}$=	**355.3908**	**348.4339**	**343.6261**					
		100	$\hat{\theta}_{min}$=	**105.1279**	**101.9131**	**99.5568**	*MEV*	ゴンペルツ･最大極値曲線			
		-0.5	$\hat{\mu}$=	**-0.5005**	**-0.3677**	**-0.2209**	*LGS*	ロジスティック曲線			
		0.5	$\hat{\sigma}$=	**0.3657**	**0.2460**	**0.3510**	*SEV*	ワイブル･最小極値曲線			
	残差平方和		S_e=	**93259.73**	**93602.41**	**95712.09**					
No.	*dose*	*x*	*y*	$\hat{y}^{MEV}$	$\hat{y}^{LGS}$	$\hat{y}^{SEV}$	*dose*	*x*	$\hat{y}^{MEV}$	$\hat{y}^{LGS}$	$\hat{y}^{SEV}$
1	0.001	-3.0	102.35	105.13	101.92	99.65	0.001	-3.00	105.13	101.92	99.65
2	0.001	-3.0	120.82	105.13	101.92	99.65	0.01	-2.00	105.13	102.24	101.09
:							0.0316	-1.50	105.13	104.36	105.85
9	0.001	-3.0	99.17	105.13	101.92	99.65		-1.25	105.23	108.55	112.22
10	0.01	-2.0	95.00	105.13	102.24	101.09	0.1	-1.00	110.10	119.43	124.69
11	0.01	-2.0	115.00	105.13	102.24	101.09		-0.75	139.74	144.93	148.05
:							0.316	-0.50	197.24	192.73	188.17
:								-0.25	256.29	254.12	246.40
63	3	0.5	234.17	339.66	341.39	343.53	1	0.00	299.16	303.28	306.24
64	10	1.0	382.00	351.29	347.49	343.63		0.25	325.22	329.93	338.30
65	10	1.0	404.32	351.29	347.49	343.63	3.16	0.50	339.66	341.39	343.53
:								0.75	347.33	345.84	343.63
71	10	1.0	366.20	351.29	347.49	343.63	10	1.00	351.29	347.49	343.63
72	10	1.0	332.67	351.29	347.49	343.63		1.50	354.34	348.31	343.63
dos e の 0.03, 0.3, 3 は，0.0316, 0.316, 3.16 の略表示								2.00	355.12	348.42	343.63

Excel のソルバーで残差平方和 S_e を最小化した結果，

MEV	ゴンペルツ・最大極値曲線	$S_e^{MEV} = 93,259.73$	-342.69
LGS	ロジスティック曲線	$S_e^{LGS} = 93,602.41$	規準
SEV	ワイブル・最小極値曲線	$S_e^{SEV} = 95,712.09$	+2109.68

に示すように，ゴンペルツ・最大極値曲線が，ロジスティック曲線に対し，残差平方和 S_e が -342.69 と減少し，あてはまりが良くなり，ワイブル・最小極値曲線は，逆に +2109.68 増加し，あてはまりが悪くなっている．

図 2 に示すのは，表の対数用量 x と子宮重量 y の散布図を描き，その上に 3 本のシグモイド曲線を重ね書きした結果である．ゴンペルツ･最大極値曲線は，$x = \log_{10}(0.10) = -1$ 過ぎてから急速に立ち上がり，$x = \log_{10}(1.0) = 0$ からは，他の曲線に比べ緩やかに上昇している．この実験は，環境ホルモンが生態系に与える影響が，どのくらいの *dose* から起きるのかを見極めて規制のための *dose* を推定することを目的としている．あてはめたシグモイド曲線から，最大反応と最小反応（$\theta_{max} - \theta_{min}$）の 10 分の 1 の推定値の 95%信頼区間の下限などが，規制用量の目安とされている．図 2 では，おおよそ $y = 120$ mg あたりで 3 本のシグモイド曲線の位置が大きく開いている．このことから，実験結果に最もあてはまるシグモイド曲線を選択する必要があり，統計的にもロジスティック曲線ではなく，ゴンペルツ・最大極値曲線が，尤もあてはまりが良いと推論される．

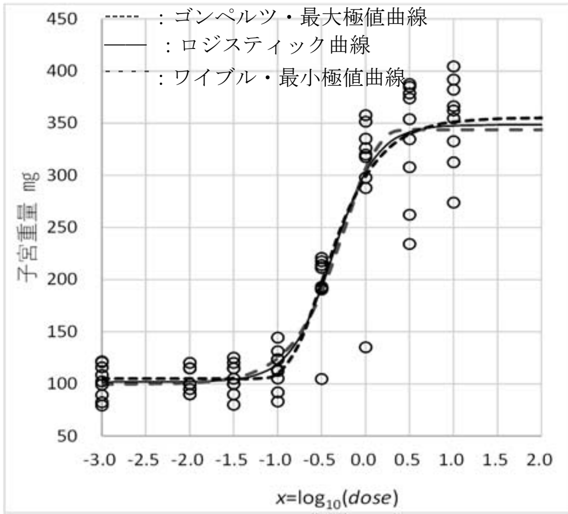


図 2　3 種のシグモイド曲線のあてはめ

4.　10 パーセント反応の用量 D_{10} の逆推定

シグモイド曲線をあてはめて最大反応の半分となる用量 $dose$ の推定は，急性毒性試験の 2 値反応データに対するプロビット法による 50%致死量　LD_{50}（*Lethal Dose* 50）の推定が，古くから行われてきた．環境ホルモンなど環境汚染物質の場合は，規制用量の設定のために 10%反応となる用量 D_{10} などの推定が求められている．詳しくは，広瀬ら（2021），「ベンチマークドース手法の適用の現状と課題 －動物実験データへの適用を中心に－」を参照のこと．

D_{10} を推定するためには，式 (2)，式 (4)，式 (7) において，D_{10} となる反応を

$$y_{D_{10}}^{分布} = \theta_{min}^{分布} + (\theta_{max}^{分布} - \theta_{min}^{分布}) \times 0.10 \tag{8}$$

と設定した場合に，$y_{D_{10}}^{分布}$ となる用量 x は，次式によって逆推定することができる．

ロジスティック曲線:
$$x_{D_{10}}^{LGS} = -\ln\left(\frac{\theta_{max}^{LGS} - \theta_{min}^{LGS}}{y_{D10}^{LGS} - \theta_{min}^{LGS}} - 1\right)\sigma_{LGS} + \mu_{LGS} \tag{9}$$

ゴンペルツ曲線:
　最大極値
$$x_{D_{10}}^{MEV} = -\ln\left(-\ln\frac{y_{D10}^{MEV} - \theta_{min}^{MEV}}{\theta_{max}^{MEV} - \theta_{min}^{MEV}}\right)\sigma_{MEV} + \mu_{MEV} \tag{10}$$

対数ワイブル曲線:
　最小極値
$$x_{D10}^{SEV} = \ln\left\{-\ln\left[-\left(\frac{y_{D10}^{SEV} - \theta_{min}^{SEV}}{\theta_{max}^{SEV} - \theta_{min}^{SEV}} - 1\right)\right]\right\}\sigma_{SEV} + \mu_{SEV} \tag{11}$$

これらの逆推定の式を用いて D_{10} を計算した結果を表 3 に示す．元の $dose$ に 10 の冪乗で戻した結果は，残差平方和の小さい順にゴンペルツ曲線が 0.156 μg/kg，ロジスティック曲線が 0.124 μg/kg，ワイブル・最小極値曲線が 0.098 μg/kg となる．この実験の目的は，環境ホルモン EE（ethinyl estradiol）の規制用量の推定なのであるが，3 種のシグモイド曲線のあてはめにより，D_{10} の推定用量が大きく異なる．図 2 から 50%反応 D_{10} ならば，違いは線の幅程度と小さいので気にすることはないが，D_{10} の場合には，残差平方和の最も小さいゴンペルツ曲線・最大極値分布の　$D_{10}^{(MEV)} = 0.156\,\mu\text{g/kg}$　を選ぶべきである．図 1 の元のデータの平均値の推移の形状からもゴンペルツ・最大極値曲線の選択の妥当性は揺るがない．

表 3　3 種のシグモイド曲線に対する 10 パーセント反応となる用量の推定

	最大	最小	位置	形状	D_{10}	D_{10}となる$\log_{10}(dose)$			$dose$
	$\theta\hat{}_{max}$	$\theta\hat{}_{min}$	$\mu\hat{}$	$\sigma\hat{}$	$y\hat{}_{D10}$	$x\hat{}^{MEV}$	$x\hat{}^{LGS}$	$x\hat{}^{SEV}$	(μ g/kg)
ゴンペルツ・最大極値 *MEV*	355.39	105.13	-0.5005	0.3657	130.15	-0.8055			0.1565
ロジスティック *LGS*	348.43	101.91	-0.3677	0.2460	126.57		-0.9082		0.1235
ワイブル・最小極値 *SEV*	343.63	99.56	-0.2209	0.3510	123.96			-1.0108	0.0975
個別の $y\hat{}_{D10} = \theta\hat{}_{min} + (\theta\hat{}_{max} - \theta\hat{}_{min})*0.10$				D_{10}=	0.10				

さて，逆推定された $\hat{D}_{10}^{分布}$ の　95%信頼区間をどのようにして求めたら良いのだろうか．最も簡便なのは，JMP の「曲線のあてはめ」を用い「Gompertz 4P」を選択し「カスタム逆推定」で　10%点に相当する $\hat{y}_{D10}^{MEV} = 130.15$　を入力すると表　4　に示すように逆推定の 95%信頼区間が得られる．

表 4　JMP の「Gompertz 4P」による逆推定

指定されたy	xの予測値	標準誤差	下側95%	上側95%
130.1500	-0.8055	0.0881	-0.9782	-0.6329

正規分布の 95%点 1.96 を使用

JMP の「非線形回帰」では，ゴンペルツ曲線式を自ら設定する必要があるが，逆推定がサポートされていて　$\hat{y}_{D10}^{MEV} = 130.15$　を入力す

表 5　JMP の「非線形回帰」による逆推定

逆推定

-Log(-Log((130.15 - θmin_M) / (θmax_M - θmin_M))) * σ_MEV + μ_MEV

指定されたy	xの予測値	標準誤差	下側0.95	上側0.95
130.1500	-0.8055	0.0881	-0.9813	-0.6298

自由度 68 の t 分布の 95%点 1.9955 を使用

ると表 5 に示すように 95%信頼区間が得られる．逆推定の計算式が表示されているが，これは，JMP 内部で設定された結果で，式 (10) に一致する．さて，95%信頼区間の計算のために必要な標準誤差の推定方法は，逆推定の式をパラメータで偏微分し，パラメータの共分散行列を用いた 2 次形式が使われている．

5. オフセットを活用したシグモイド曲線に対する逆推定

10 パーセントの反応となるような用量 x を逆推定し, その 95%信頼区間を推定したいとの目的のためには, 逆推定の機能が組込まれている JMP が最強である．ただし，有償であり誰でも手軽に使えるわけではない．そこで，無償で継続的に使える OnDemand SAS の NLIN プロシジャにより，オフセットを活用した逆推定値の 95%信頼区間を導入し，さらに，Excel により推定結果の確認を行なう．

累積分布関数を活用したシグモイド曲線のあてはめのパラメータとして $\theta_{max}^{分布}$，$\theta_{min}^{分布}$，$\mu_{分布}$，$\sigma_{分布}$ の 4 つのパラメータを用いてきた．シグモイド曲線について，位置パラメータと形状パラメータを用いて次のような規準化

$$\eta^{LGS} = \frac{x-\mu_{LGS}}{\sigma_{LGS}}, \qquad \eta^{MEV} = \frac{x-\mu_{MEV}}{\sigma_{MEV}}, \qquad \eta^{SEV} = \frac{x-\mu_{SEV}}{\sigma_{SEV}} \tag{12}$$

を行ってきた．この基準化した変数 $\eta^{分布}$ を用いることにより，$\theta_{max}^{分布}=1$，$\theta_{min}^{分布}=0$ とするシグモイド曲線は,

$$\pi^{LGS} = \frac{1}{1+\exp(-\eta^{LGS})}, \qquad \pi^{MEV} = \exp[-\exp(-\eta^{MEV})], \qquad \pi^{SEV} = 1-\exp[-\exp(\eta^{SEV})] \tag{13}$$

のように簡単化することができる．基準化した変数 $\eta^{分布}$ が 0.0 となるのは，それぞれ（$\hat{\mu}_{LGS}$，$\hat{\mu}_{MEV}$，$\hat{\mu}_{SEV}$）であることは式 (12) から自明である．

これらのシグモイド曲線のパーセント点 $\pi^{分布}$ を $\eta^{分布}$ に関して解くと

$$\eta_{\pi}^{LGS} = \ln\left(\frac{\pi^{LGS}}{1-\pi^{LGS}}\right), \qquad \eta_{\pi}^{MEV} = -\ln[-\ln(\pi^{MEV})], \qquad \eta_{\pi}^{SEV} = \ln[-\ln(1-\pi^{SEV})] \tag{14}$$

が得られる．表 6 に示すように，これらの変数 $\eta_{\pi}^{分布}$ が，任意の反応のパーセント点 $\pi^{分布}$ （0.0～1.0）に対するシグモイド曲線の X 軸の位置となっている．これらの変数 $\eta_{\pi}^{分布}$ を式 (12) に加えることにより，推定される位置パラメータ $\mu_{分布}$ が元の位置からオフセットされたパラメータ $\mu_{分布}^{offset}$ として推定される．

実際に反応が $\pi=0.10$ となるオフセット変数 $\eta_{\pi=0.10}^{分布}$ を式 (4) に加えたゴンペルツ曲線は,

$$\pi^{MEV} = \exp\left[-\exp\left(-\frac{x-\mu_{MEV}^{offset}}{\sigma_{MEV}} + \eta_{\pi=0.10}^{MEV}\right)\right] \tag{15}$$

のように設定することができる．この式で，推定された位置パラメータ $\mu_{分布}^{offset}$ が，どうして 10 パーセント反応量となるか，謎めいており，なかなか理解しづらい．実際のデータ y_i に対しシグモイド曲線のあてはめる場合を想定する．残差線形化法によりシグモイド曲線の推定値 $\hat{y}_i$ に対する位置パラメータ $\hat{\mu}_{分布}$ が得られたとしたときに，オフセット変数が $\eta_{\pi}^{分布}=0.0$ となっていれば，なにも悪さはしないので通常の位置パラメータが推定される．シグモイド曲線の 10 パーセント点は $\eta_{\pi=0.10}^{分布} \neq 0.0$ なので，実際のデータ y_i にシグモイド曲線をあてはめようとしたときに，$\eta_{\pi=0.10}^{分布}$ が邪魔をするので，本来の位置パラメータ $\hat{\mu}_{分布}$ ではなく，邪魔された分だけ位置を変えて $\hat{\mu}_{分布}^{offset}$ が，残差平方和を最小にするパラメータとして推定される．したがって，オフセット変数 $\eta_{\pi}^{分布}$ を適切に調

表 6　各シグモイド曲線に対応するオフセット変数 $\eta_{\pi}^{分布}$ の推定値

パーセント点		η_{π}^{LGS}	η_{π}^{MEV}	η_{π}^{SEV}
π	変曲点	=ln (π/(1−π))	=−ln (−ln (π))	=ln (−ln(1−π))
0.9000		2.1972	2.2504	0.8340
0.6321	=exp(−exp(−0))	0.5413	0.7794	0
0.5000	=1/(1+exp(−0))	0	0.3665	-0.3665
0.3679	=1−exp(−exp(0))	-0.5413	0	-0.7794
0.1000		-2.1972	-0.8340	-2.2504

整することにより，任意のパーセント点の推定が可能となる．

シグモイド曲線のパーセント点についての逆推定を行なうためのオフセット変数$\eta_{\pi}^{分布}$がどのように変化するかを概観する．表6に示したのは，幾つかのパーセント点に対し，各シグモイド曲線のオフセット値を計算した結果である．図3に示すのは，ゴンペルツ・最大極値曲線π^{MEV}に対し表6に示した各パーセント点に対するオフセット値を重ね書きした結果である．これだけを見れば，単にゴンペルツ・最大極値曲線のパーセント点π^{MEV}に対するη^{MEV}の位置を示しているに過ぎない．ただし，オフセット値を含む式 (15) を用いて実際のデータに対し残差線形化法（ガウス・ニュートン法）により推定した場合に，位置パラメータ$\hat{\mu}_{MEV}^{offset}$の95%信頼区間を得ることができる．

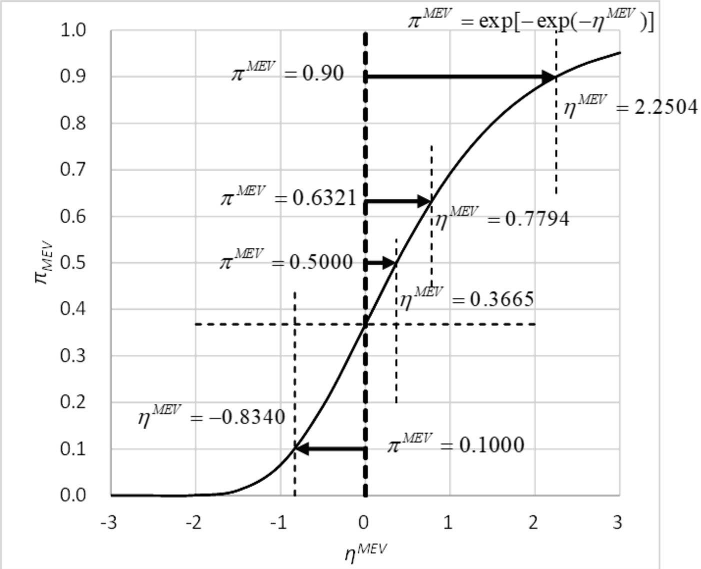


図3　ゴンペルツ・最大極値曲線に対するオフセット

6. SAS の非線形 NLIN プロシジャによる解析

オフセットを含む解析に先立ち，NLIN プロシジャによるゴンペルツ・最大極値曲線のあてはめを行いパラメータの推定結果を, Excel の結果と照合する．NLN プロシジャは，初期値設定のための parms ステートメントでの初期値設定に引き続き，model ステートメントでの式 (4) で示した式を設定する．

```
TITLE1  '環境ホルモン.SAS  << ラットの子宮重量  >> ' ;
DATA   d01  ;
   input  dose @ ;
   x = log10(dose) ;
   do i=1 to0 9 ;
     input  y  @ ; output ;
   end;
/* dose    1      2      3      7      8      9     11     18     19 */
datalines ;
0.0010  102.35 120.82 115.92 121.62  79.22 108.47  82.45  89.25  99.17
0.0100   95     115    115    120     90    115    100     90    100
0.0316  105     115    120    125     80    115    100     90    100
 0.100  112.22 123.47 144.42 131.25 105.08 123.60 113.38  91.80  83.17
 0.316  190.45 217.48 213.95 220.83 211.13 211.37 191.23 193.07 104.67
  1.00  319.78 351.32 326.07 317.52 287.68 357.57 297.67 334.95 135.17
  3.16  373.72 384.72 378.37 387.43 262.20 353.82 307.60 334.48 234.17
 10.00  382.00 404.32 354.37 391.67 273.73 362.05 312.40 366.20 332.67
;
proc nlin data=d01  list ;
    parms     θmaxM=350  θminM=100  μ_MEV=-0.5   σ_MEV=0.5 ;
           g = Exp( -(x-μ_MEV)/σ_MEV ) ;
    model  y = θminM + ( θmaxM - θminM ) * Exp( - g );
run ;
```

表7にゴンペルツ・最大極値曲線のパラメータの推定結果を示す．残差平方和$S_e = 93259.7$と表2に示した Excel での結果に一致することが確認できる．

オフセット無しでの結果の再現性が確認できたので, 反応の10パーセント点における逆推定を式 (13)中，および，式 (14)中 で示したオフセット変数η_{π}^{MEV}を用いて NLN プロシジャに組み込み推定する．NLIN プロシジャの中で「π=0.1」を与え，式 (15) により，model 式の中ので「(-log(-log(π)))」として与えている．ここで,「π=0.9」とすれば，90パーセント点を推定する

表7　ゴンペルツ・最大極値曲線のパラメータの推定

要因	自由度	平方和	平均平方	F 値	近似Pr > F
Model	3	785681	261894.00	190.96	<.0001
Error	68	93259.7	1371.50		
Corrected	71	878941			
パラメータ	推定値	標準誤差	近似 95% 信頼限界		（注）
θmaxM	355.4	13.5461	328.4	382.4	t 分布の
θminM	105.1	6.6476	91.8628	118.4	両側5%点
μ_MEV	-0.5005	0.0534	-0.6071	-0.3938	
σ_MEV	0.3657	0.0881	0.1900	0.5415	

ことができる．

表 8 に示すようにパラメータ「μ_MEV_offset」の行に 10 パーセント点の推定値 $\hat{x}^{MEV}_{y=D_{10}} = -0.8055$ が得られ，近似標準誤差が $SE(\hat{x}^{MEV}_{y=D_{10}}) = 0.0931$ となる．この結果は，表 5 に示した JMP の逆推定で求めた $SE = 0.0881$ とは異なる．JMP では，逆推定式をパラメータに関した偏微分した式を用いたデルタ法での近似計算であるのに対し，NLIN プロシジャでは，非線形モデルとして直接推定しているためである．表 10 に示すように Excel の計算シートにオフセットを加えれば，SAS の NLIN プロシジャと同じ結果が得られる．

```
proc nlin data=d01  ;
    π=0.1 ;
    offset = -log(-log(π)) ;
    parms      θmaxM=350  θminM=100  μ_MEV_offset=-0.5   σ_MEV=0.5 ;
            g = exp(- ( (x - μ_MEV_offset)/σ_MEV + offset )) ;
    model  y = θminM + ( θmaxM - θminM ) * exp( - g );
run ;
```

表 8　ゴンペルツ・最大極値曲線の 10%点に対する逆推定

要因	自由度	平方和	平均平方	F 値	近似Pr > F
Model	3	785681	261894	190.96	<.0001
Error	68	93259.7	1371.5		
Corrected Total	71	878941			

パラメータ	推定値	標準誤差	近似 95% 信頼限界	
θmaxM	355.4	13.5460	328.4	382.4
θminM	105.1	6.6476	91.8629	118.4
μ_MEV_offset	-0.8055	0.0931	-0.9913	-0.6197
σ_MEV	0.3657	0.0881	0.19	0.5415

```
data d02 ;    /* 95% 信頼区間 */
    retain dose ;
    do  x = -3, -2 ;              dose=10**x;  output; end;
    do  x = -1.5 to 1 by 0.25 ;  dose=10**x;  output; end;
    do  x =  1.5, 2 ;             dose=10**x;  output; end;

data d03 ;
    set d01 d02 ;

proc nlin data=d03  ;
     parms     θmaxM=350   θminM=100   μ_MEV=-0.5   σ_MEV=0.5 ;
           g = exp(-(x-μ_MEV)/σ_MEV ) ;
    model  y = θminM + (θmaxM-θminM)*exp( - g );
     output out=Out03 predicted=y_hat STDP=SE L95M=L95 U95M=U95 ;
run ;
proc print data=out03 ; run;
```

7.　95%信頼区間

NLIN プロシジャで計算された 10 パーセント点の推定値 $\hat{x}^{MEV}_{y=D_{10}} = -0.8055$，および，95%信頼区間 (-0.9913，-0.6197) が適切に推定されているか自己検証するには，どうしたら良いのであろうか．JMP のように手軽に結果のグラフ表示ができれば良いのだが，SAS ではなかなか思うようにできないので，Excel の力を借るのが現実的である．そのために，解析用の SAS データセットに，新たな推定用のデータを結合する．

表 9　ゴンペルツ・最大極値曲線の推定値および 95%信頼区間の計算結果

OBS	dose	x	i	y	y_hat	SE	L95	U95
1	0.001	-3	1	102.35	105.13	6.6476	91.86	118.39
:								
72	10	1	9	332.67	351.29	10.445	330.45	372.13

OBS	dose	x	i	y	y_hat	SE	L95	U95
73	0.001	-3	.	.	105.13	6.648	91.86	118.39
74	0.01	-2	.	.	105.13	6.648	91.86	118.39
75	0.032	-1.5	.	.	105.13	6.647	91.86	118.39
76	0.056	-1.25	.	.	105.23	6.505	92.26	118.21
77	0.1	-1	.	.	110.10	7.706	94.72	125.48
78	0.178	-0.75	.	.	139.74	14.654	110.50	168.98
79	0.316	-0.5	.	.	197.32	11.563	174.24	220.39
80	0.562	-0.25	.	.	256.29	11.027	234.29	278.30
81	1	0	.	.	299.16	10.805	277.60	320.72
82	1.778	0.25	.	.	325.22	8.478	308.30	342.14
83	3.162	0.5	.	.	339.68	7.612	324.49	354.87
84	5.623	0.75	.	.	347.33	8.848	329.68	364.99
85	10	1	.	.	351.29	10.445	330.45	372.13
86	31.62	1.5	.	.	354.34	12.441	329.51	379.17
87	100	2	.	.	355.12	13.193	328.80	381.45

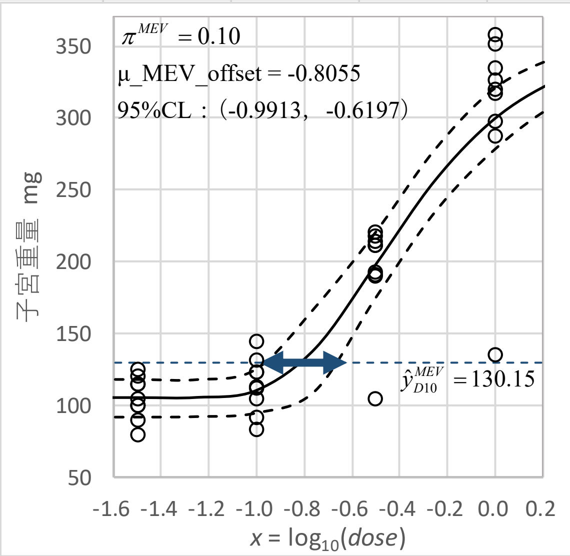


図 4　SAS の出力を Excel で作図し 10%点の 95%信頼区間の表示

図 4 は，表 9 に示した SAS の出力を Excel に取り込んだデータに基づいて作成している．10 パーセント反応量における逆推定は，表 8 のパラメータ μ_offsetM についての「近似 95%信頼限界」が重ね書きされている．よく見ると曲線の 95%信頼区間の幅よりも広めになっているが，計算方法に依存し微妙な差異となるためである．

8. 考察

化学物質の生物に対する様々な反応は，対数用量（濃度）に対しシグモイド曲線状になることが経験的に知られている．古くは，急性毒性試験における 50%が死亡する用量 LD_{50} を推定し，その 95%信頼区間を示すことが当然のこととして受け入れられ，そのための推定法としてプロビット法が普及し，SAS の初期のバージョンから PROBIT プロシジャが提供されていた．無償で継続的に使用できる OnDemand SAS でも，もちろんバージョンアップ版が使えるようになっている．しかるに，その計算原理は難解であり，きちっと解説されている日本語の成書は見当たらない．詳しくは，高橋（2017），「一般化線形モデルを Excel で極め活用する－プロビット法・ロジット法・補 2 重対数法－」を参照のこと．

酵素反応速度論におけるミカエリス・メンテン式は，量的な反応に対し最大反応となるパラメータとしての最大速度 *Vmax* , 基質濃度 [*S*] の 50%反応となるパラメータ *Km*（ミカエリス定数）によって定義される．プロビット法による LD_{50} と同様のパラメータであるので，「ミカエリス定数 *Km* の 95%信頼区間」について Web で検索しても全くヒットしない．ミカエリス・メンテン式は，非線形モデルの代表的な事例であり，SAS の NLIN プロシジャのユーザズ・ガイドの最初の事例としても使われており，もちろんミカエリス定数 *Km* の 95%信頼区間も例示されている．得られたパラメータについて 95%信頼区間を付けるのは，最低限のマナーであると認識しているのであるが, 推定されたミカエリス定数 *Km* については, 95%信頼区間を付けることが, まったく習慣化されていないと認識を新たにした．

他の応用分野でも，伝統的な統計解析手順による入門書が数多く出版され，その解析結果を得るための統計ソフトの使い方が解説されるのが常であり，統計ソフトが新たに切り開いた解析方法については，まったく置き去りにされるのが現状である．典型例は，SAS の GLM プロシジャにより 40 年以上前に提供された“最小 2 票平均，LSMEANS”である．SAS ユーザーにとっては，あたりまえの統計量であるが，「どのように計算しているのか，その 95%信頼区間の計算方法は，いかに」に答えられるのであろうか．詳細は，高橋行雄（2020），「最小 2 乗平均の謎を予測プロファイルで解く」を参照のこと．

シグモイド曲線のあてはめも伝統的には，線形変換を工夫し重み付き回帰で解析する方法が，田栗正章・番場弘・浅井晃（1973），「重みつき最小 2 乗法による回帰曲線の当てはめ」などによって研究されていた．古い時代の非線形モデルのための SAS の NLIN プロシジャは，非線形式のみならずすべてのパラメータの偏微分式も指定しなければならず，手軽に使える状況ではなかった．しかし，現在では，自動微分の機能が追加され非線形式とパラメータの初期値の設定のみで手軽に使えるようになっている．無償で継続的に使用できる OnDemand SAS でも NLIN プロシジャのフルセットが使える状況であるが，その存在は影が薄いので，実際のデータを用いた解析事例を充実し，Excel のよる計算方法示す活動が必要である．

他方，統計ソフトの計算方法が，ユーザズ・ガイドに示されているとは言え，難解でブラック・ボックス的であり，普及の妨げとなる．芳賀敏郎（2016），「医薬品開発のための統計解析，第 3 部 非線形モデル 改訂版」では，Excel のソルバーと JMP の「非線形回帰」を相互に補完的に使うことにより，非線形モデルに関する優れた入門書である．ただし，非線形回帰で推定されたパラメータの標準誤差の計算方法については，割愛されている．Excel のソルバーでは，表 2 に示したように，パラメータの推定値が得られるだけで，パラ

メータの標準誤差が得られない．したがって，推定されたパラメータの 95%信頼区間を求めることができない．どのような計算法によって標準誤差が得られるのだろうか.

限られたページの中で Excel による計算法を示すことができなかったが，表 10 に示すように，式 (4) のゴンペルツ曲線式を 4 つのパラメータで偏微分し，72×4 の微係数行列 $\boldsymbol{Z}$ を求め，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\xi}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\hat{\sigma}^2$ の対角要素がパラメータの分散となることから，その平方根が標準誤差 SE となる．求められた SE から 95%信頼区間が（–0.6071，–0.3938）と推定され，表 7 に示した NLIN プロシジャの結果に一致する.

表 10　Excel によるゴンペルツ・最大極値曲線の推定値の 95%信頼区間の計算結果

ξ	ξ^$^{(m-1)}$		$\Sigma(\xi$^$)=(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\sigma$^^{2}					SE	L95%	U95%
θ^$_{max}$=	**355.392**		**183.50**	-13.02	0.220	0.887		13.546	328.36	382.42
θ^$_{min}$ =	**105.128**		-13.02	**44.190**	0.134	-0.153		6.6476	91.86	118.39
μ^$_{MEV}$=	**-0.5005**		0.220	0.134	**0.003**	0.000		0.0534	-0.6071	-0.3938
σ^$_{MEV}$=	**0.3657**	σ^^{2}	0.887	-0.153	0.000	**0.008**		0.0881	0.1900	0.5415
S_e =	**93259.73**	**1371.47**					$t_{0.05}$(72-4)=	1.9955		

シグモイド曲線状となる反応データに対する統計解析の普及のために，活動を継続的に続けている．その一環として，今回は，逆推定をテーマとし，(1) Excel のソルバーによるパラメータの推定，(2) JMP の「曲線のあてはめ」および「非線形回帰」による手軽な逆推定の例示，(3) オフセットを用いた SAS/NLIN プロシジャによる逆推定，(4) シグモイド曲線の 95%信頼区間の例示，(5) Excel の行列関数を活用したパラメータの 95%信頼区間の計算法，について示した．更なる普及のために，データに基づく解析事例を掲載した本の出版が必要不可欠と認識している．現在，出版準備中の「層別因子を含む探索的な回帰分析入門」の「第 11 章 各種のシグモイド曲線を用いた逆推定」の推敲のために本発表を行った.

参 考 文 献

1） 山田雅之, 吉田光宏, 高橋行雄(2003), 陰性および陽性対照があるシグモイド曲線－ダミー変数を持つ非線型回帰モデルの応用－, SAS ユーザー総会論文集;51-60.
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/sugj2003.pdf

2） 広瀬明彦, 西浦博(2021), ベンチマークドース手法の適用の現状と課題 －動物実験データへの適用を中心に－, 産業医学レビュー, Vol.34, No.1;1-15. https://www.jstage.jst.go.jp/article/ohpfrev/34/1/34_1/_pdf/-char/ja

3） 高橋行雄(2017), 一般化線形モデルを Excel で極め活用する－プロビット法・ロジット法・補 2 重対数法－, 第 6 回続高橋セミナー, https://www.yukms.com/biostat/takahasi2/rec/006.htm

4） 高橋行雄(2020), 最小 2 乗平均の謎を予測プロファイルで解く, 第 9 回続高橋セミナー,
https://www.yukms.com/biostat/takahasi2/rec/009-13.htm

5） 高橋行雄(2021), 最尤法によるポアソン回帰入門, 第 13 章 最小 2 乗平均の謎を予測プロファイルで解く;421-460, カクワークス社.

6） 田栗正章, 番場弘, 浅井晃(1973), 重みつき最小 2 乗法による回帰曲線の当てはめ, 応用統計学, Vol.2, No.2; 95-112.

7） SAS Institute(2021), SAS/STAT13.2 User’s Guide, The NLIN Procedure,
https://support.sas.com/documentation/onlinedoc/stat/132/nlin.pdf

8） 芳賀敏郎(2016), 医薬品開発のための統計解析, 第 3 部 非線形モデル 改訂版, サイエンティスト社.

SAS初心者のためのDATAステップ処理・データセットマージ入門

○雨宮　祐輔
（第一三共株式会社　データインテリジェンス部）

Introduction to Data Step Processing and Dataset Merge for SAS Programing Beginner

Yusuke Amemiya
Data Intelligence Department, Daiichi Sankyo Co., Ltd.

要旨

SAS プログミングの初心者にとって DATA ステップ処理ならびにデータマージは理解し難く、表面上不可視の部分が多く存在する。本稿は DATA ステップ処理として PDV、コンパイルフェーズ、実行フェーズを説明したうえで、データマージにおける注意事例を紹介する。

キーワード：DATA ステップ処理、プログラムデータベクトル、PDV、MERGE ステートメント

1　はじめに

DATA ステップでのデータセット作成において、可視化されないプログラムデータベクトル（以下、PDV）の挙動、コンパイルフェーズおよび実行フェーズは log の確認やデータマージを理解するために非常に重要である。しかしながら SAS 初心者にとって、これらの処理を理論的に理解することは難しい場合が多い。本稿は、SAS のデータ処理を理解するための一助となるように SAS データセットを読み込んだ際の挙動を図示、説明する。また、データマージにおける注意事例を紹介し、事例の発生理由をデータ処理の観点から説明する。まずは図を確認して概要をつかんでから本文を読んでいただきたい。なお、SAS プログラムは SAS OnDemand for Academics（バージョン 9.4 M7 (確認日: 2023/08/30) ）で実行した。

2　DATA ステップ処理

2.1 処理の概略

ここで説明する DATA ステップ処理は DATA ステップのコードをサブミット（実行命令）した結果、SAS データセットを読み込み、新規の SAS データセットを作成する処理を指す。この処理の概略は次頁に示している（図 1）。処理の流れとして、(1) プログラムを実行する前にプログラムの構文（使用方法）や作成する変数に誤りがないか等の確認を行うコンパイルフェーズ、(2) プログラムに問題がないことを確認した後に、プログラムを実行してデータセットの作成を行う実行フェーズ、といった 2 つのフェーズが実行されている。まずは、処理が 2 種類あること、データセット作成には PDV が用いられることを押さえてほしい。

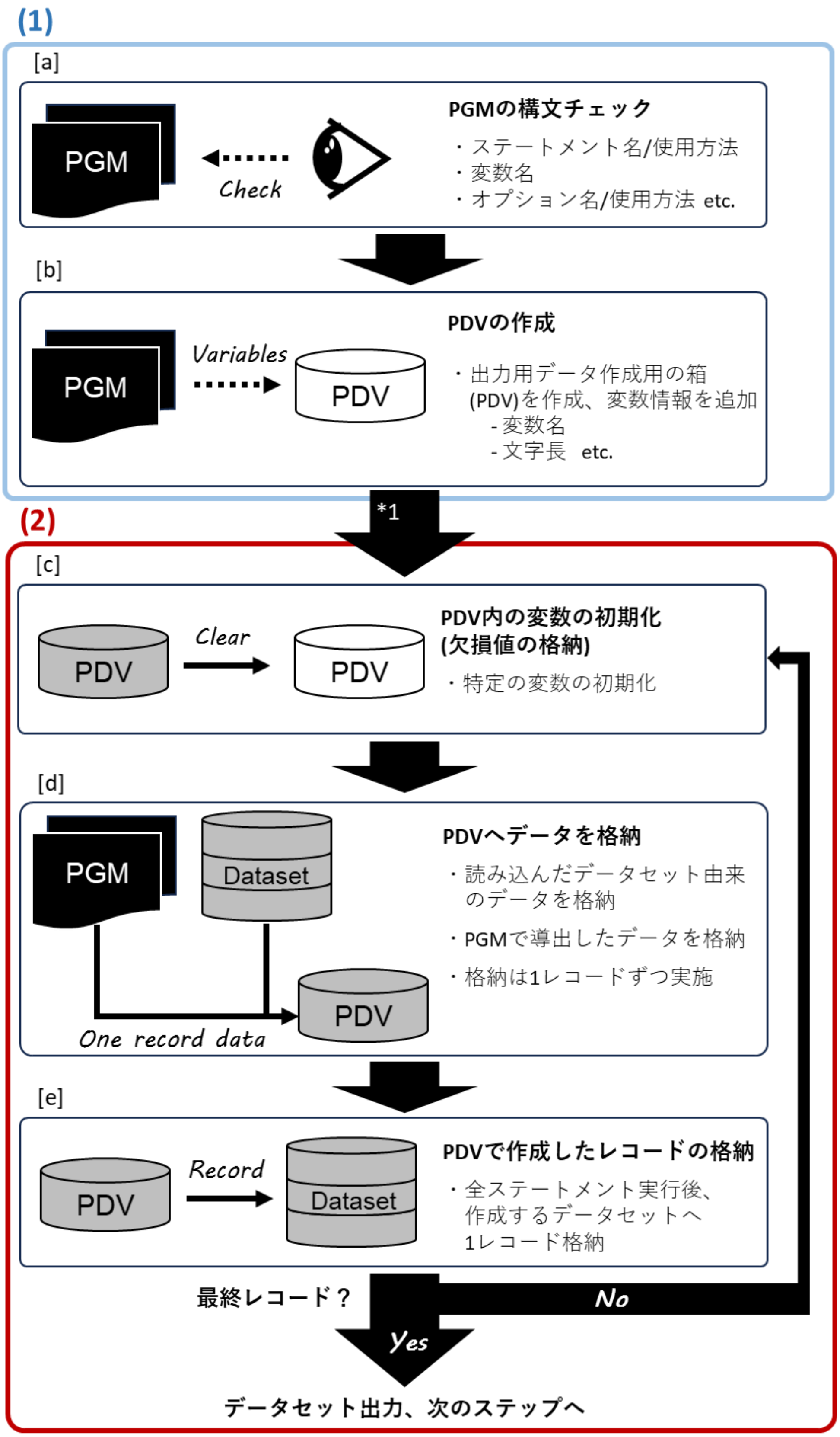


図 1. DATA ステップ処理の流れ

(1)-水色部分: コンパイルフェーズ

(2)-赤色部分: 実行フェーズ

*1 データセットのディスクリプタ部（データセット名やレコード数などのデータセット関連の情報や変数名や属性の情報）作成が実行されているが、図での説明を省略した。

2.2 PDV (プログラムデータベクトル)

PDV は、コンパイルフェーズおよび実行フェーズで重要な役目を果たす「データセットを作成するためのデータ用の入れ物（変数）を保有する作業用バッファー」である（図 2）。

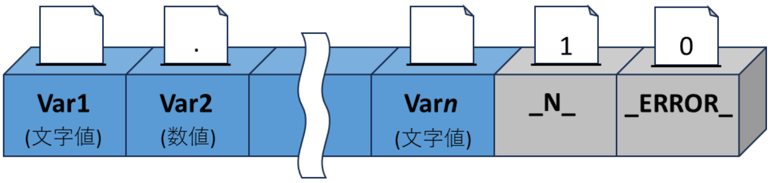


図 2. PDV のイメージ
1 レコード分のデータを格納するための入れ物（変数）が用意されている。実行時に数値や文字値が格納される。図は実行フェーズ開始時における値の格納イメージ。
青色：プログラムならびに入力用のデータセットで定義された変数（Var1, Var2, Var*n* は変数名を表す）。
灰色：データセットに出力されないが PDV 上でのみ作成される変数（詳細は「実行フェーズ」で説明）。

PDV は、コンパイルフェーズでは変数名などの変数情報を準備し、実行フェーズでは 1 レコードずつ値を格納してデータセットに出力させている。
また、PDV では 2 種類の変数が存在する。1 つは、プログラムもしくは読み込んだデータセット由来の変数である。2 つ目は、データ処理のために一時的に PDV 上でのみ作成される_N_変数や_ERROR_変数等の自動変数である。_N_変数はデータステップの処理回数をカウントし、_ERROR_変数はデータステップ実行中にエラーが発生したか否かを示す。これらの変数はレコードが正しく作成されるうえで重要な変数であり、実行フェーズで値が格納される。
DATA ステップ処理は PDV を作業場として、データセットの作成を行っている。以降で PDV を使用した処理としてコンパイルフェーズと実行フェーズで何が実施されているのか確認する。

2.3 コンパイルフェーズ

コンパイルフェーズは、プログラム処理開始前の準備段階であり、データセットの作成前に SAS の構文として誤ったプログラムを実行しないように点検を行っている。このフェーズでは実行するプログラムの構文チェックを行ったうえで、PDV に変数情報を追加する処理が行われている（図 3）。

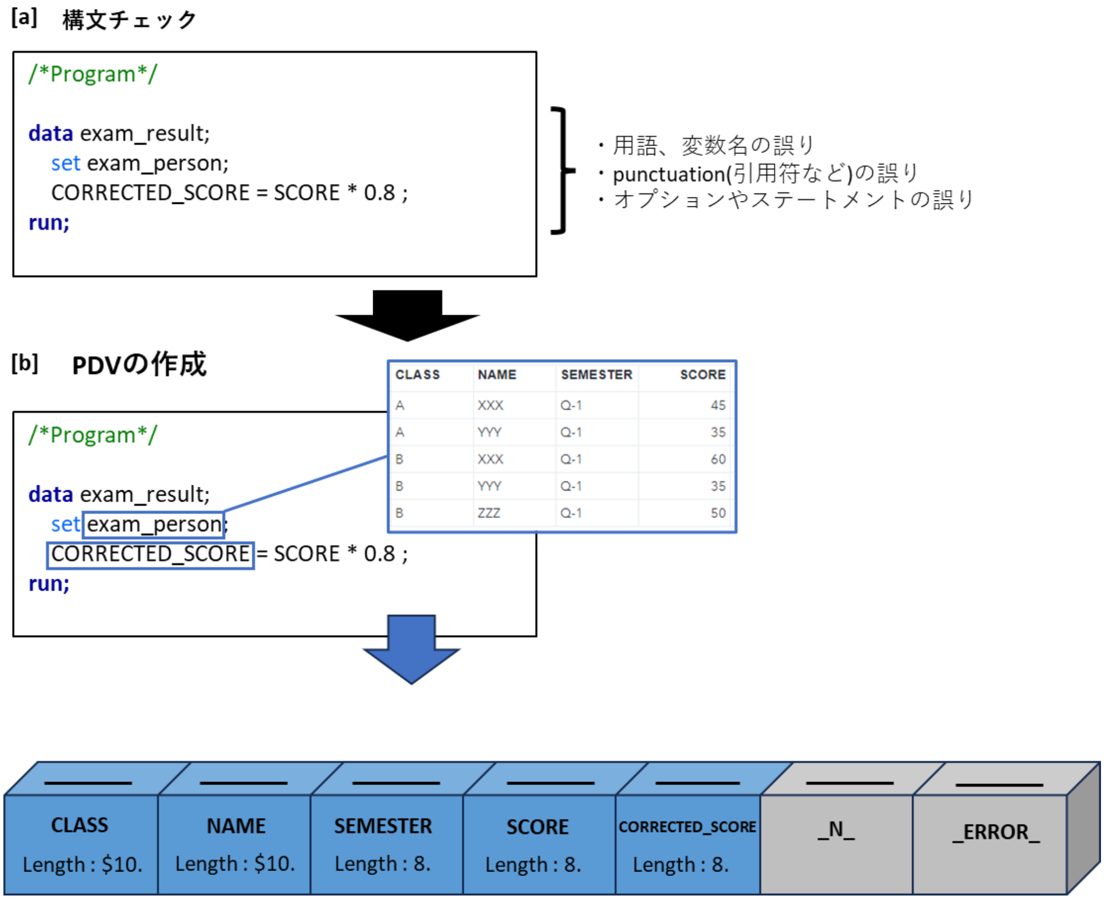


図 3. コンパイルフェーズの流れ[*1]

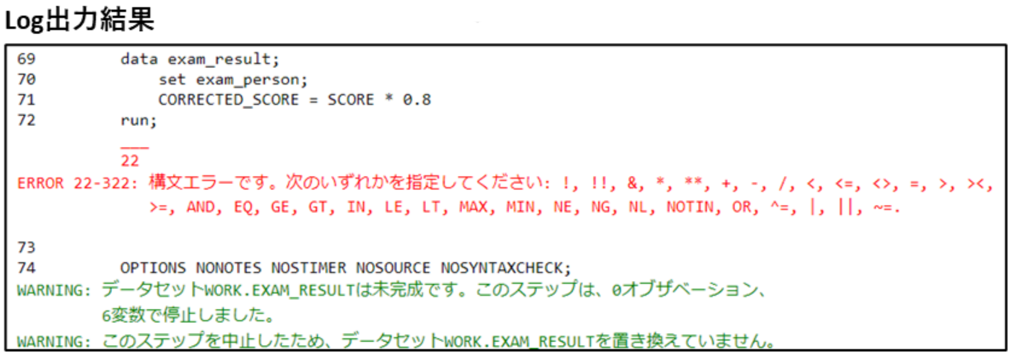
[a] 構文チェックにおけるチェック対象例

[b] PDV への変数情報付与のイメージ (図 2 のイメージと異なりこのフェーズでは値は格納されていない)

*1 本来ならデータセットの情報付与（ディスクリプタ部の作製）も実施されるが説明を省略している

手順の詳細としては [a] プログラムにおける誤りの有無を確認（構文チェック）して問題なければ、 [b] DATA ステップ処理で発生する「変数の情報」を付与する、という 2 つの手順が実行されている。ここで記載する「変数の情報」は、SET ステートメントで指定したデータセット内に含まれる変数（複数データセットを読み込む場合は先に読み込まれた変数情報を採用する）、ならびにプログラム上で定義した新規作成変数の 2 つを指している。また、変数に値は格納されない。

また、[a]の構文チェックで error を確認した場合、log に error 内容を出力し、処理を停止する（図 4）。



Log出力結果

```
69          data exam_result;
70              set exam_person;
71              CORRECTED_SCORE = SCORE * 0.8
72          run;
            ___
            22
ERROR 22-322: 構文エラーです。次のいずれかを指定してください: !, !!, &, *, **, +, -, /, <, <=, <>, =, >, ><,
              >=, AND, EQ, GE, GT, IN, LE, LT, MAX, MIN, NE, NG, NL, NOTIN, OR, ^=, |, ||, ~=.

73
74          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
WARNING: データセットWORK.EXAM_RESULTは未完成です。このステップは、0オブザベーション、
         6変数で停止しました。
WARNING: このステップを中止したため、データセットWORK.EXAM_RESULTを置き換えていません。
```

図 4 構文エラーと log 出力例（セミコロンが欠落しているプログラムを実行した場合 *1）

*1 セミコロンの記載漏れの場合には、セミコロン記載漏れの指摘ではなく、ステートメントや導出における構文の誤りを error として指摘する

また、このフェーズではあくまで変数情報を付与するだけであり、PDV には値が格納されない。変数の格納値に由来する error または warning については、レコード作成を行う実行フェーズで確認される。

2.4 実行フェーズ

実行フェーズでは、構文チェックで問題ないことを確認したプログラムを処理し、格納値に関するエラーを確認の上でデータセットの作成を行う。データセットの作成は全レコードをまとめて処理するのではなく、レコードを 1 つずつ処理してデータセットに出力している。また、このフェーズでは_N_変数および_ERROR_変数を含め、PDV の各変数に値が格納される。処理の手順は図 5 に示す通りである。

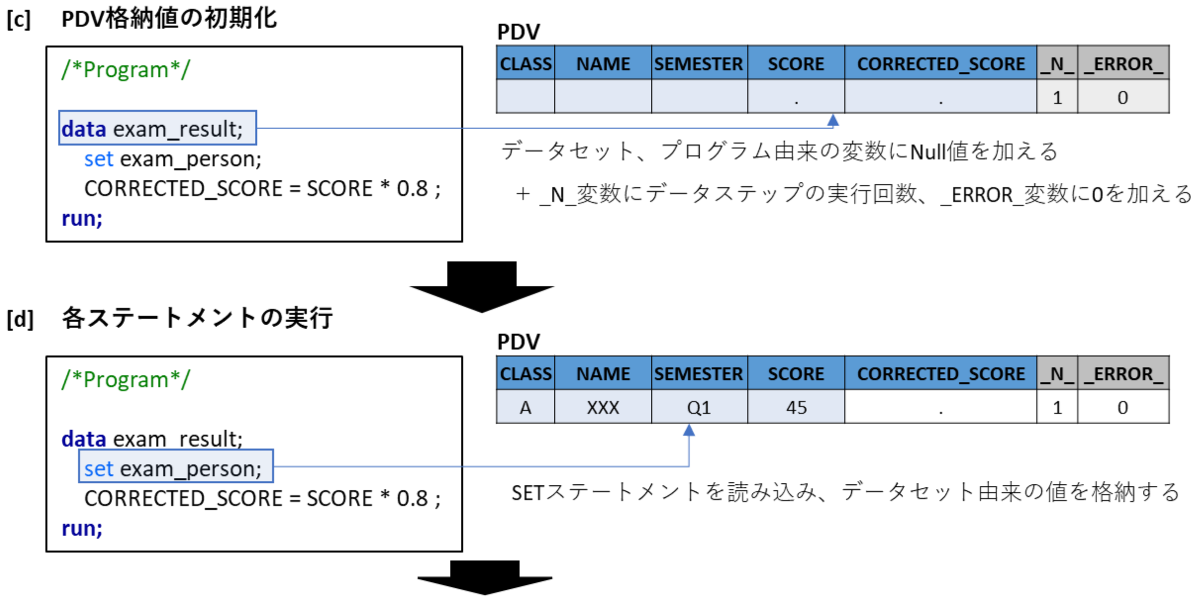


図 5. 実行フェーズの流れ

実行フェーズにおけるプログラムの実行イメージと PDV への値格納イメージ。図中で実行されたプログラムならびに PDV において値の格納が行われている部分は塗りつぶしを行っている。

[c], [c’] PDV 格納値の初期化イメージ（c, c’で処理内容が若干異なるため差別化）

[d] 各ステートメントの実行イメージ

[e] 作成したレコードの出力用データセットへ格納（OUTPUT ステートメントが存在しない場合）

実行フェーズの処理を順番に説明すると、まずは、 [c] DATA ステートメントを読み込み、PDV の初期化を行う。PDV に存在する変数のうち格納するデータセットならびにプログラム由来の変数へ Null 値を格納する。PDV 特有の変数である_N_には数値の 1 を、_ERROR_には数値の 0 を格納する。

続いて、[d] DATA ステップ内のステートメントを処理する。図 5 のプログラムの場合、まずは SET ステートメントを読み込み、sas データセットの 1 レコード目に由来するレコードの値を PDV の各変数に格納する。

値が存在しない場合は、変数に値が格納されずに PDV は Null 値が格納された状態となる。続いて「CORRECTED_SCORE = SCORE * 0.8」を実行し、結果を変数に格納する。

そして、[e] RUN ステートメントを読み込み、出力を命令するステートメント（OUTPUT ステートメント）が DATA ステップ内に存在しなければ、PDV に格納されている値を出力用のデータセットに格納する（暗黙の OUTPUT とも呼ぶ）。

RUN ステートメントを読み込んだ後は、[c'] 最初に読み込んだ DATA ステートメントに戻ったうえで PDV を初期化する。この時、データステップの反復回数が 2 回目であることを PDV に記録するために_N_に 1 を足し、_ERROR_変数には 0 を格納する。そして、すべてのレコードの処理が完了していない場合には、変数の初期化（Null 値の格納）を行う。初期化の対象となる変数は、SET ステートメントで指定したデータセットのレコードではなく、プログラムで割り当てた変数である。2 行目以降のレコードを読み込んだ際に PDV を初期化する際は、すべての変数を対象としていないことに注意いただきたい。

[c']から[e]の手順を繰り返し実行し、すべてのレコード処理が完了した場合には、出力用データセットに記録していたレコード群を DATA ステップで指定したデータセットに出力する、というのが実行フェーズの役割である。

ただし、入力された値もしくは読み込んだデータセット等にふさわしくない値があるなど構文チェックで確認した以外のエラーが発生した場合には、PDV にて_ERROR_変数に数値の 1 が格納され、log に実行結果を出力する（図 6）。

```
/*Program*/

data exam_result;
   set exam_person;
   CORRECTED_SCORE = SEMESTER * 0.8 ;
run;
```

LOGウィンドウ

```
70         data exam_result;
71             set exam_person;
72             CORRECTED_SCORE = SEMESTER * 0.8;
73         run;

NOTE: 以下の箇所で文字値を数値に変換しました。(行:カラム)
      72:23
NOTE: 無効な数値データSEMESTER='Q-1'が行 72 カラム 23にあります。
CLASS=A NAME=XXX SEMESTER=Q-1 SCORE=45 CORRECTED_SCORE=. _ERROR_=1 _N_=1
NOTE: 無効な数値データSEMESTER='Q-1'が行 72 カラム 23にあります。
CLASS=A NAME=YYY SEMESTER=Q-1 SCORE=35 CORRECTED_SCORE=. _ERROR_=1 _N_=2
NOTE: 無効な数値データSEMESTER='Q-1'が行 72 カラム 23にあります。
CLASS=B NAME=XXX SEMESTER=Q-1 SCORE=60 CORRECTED_SCORE=. _ERROR_=1 _N_=3
NOTE: 無効な数値データSEMESTER='Q-1'が行 72 カラム 23にあります。
CLASS=B NAME=YYY SEMESTER=Q-1 SCORE=35 CORRECTED_SCORE=. _ERROR_=1 _N_=4
NOTE: 無効な数値データSEMESTER='Q-1'が行 72 カラム 23にあります。
CLASS=B NAME=ZZZ SEMESTER=Q-1 SCORE=50 CORRECTED_SCORE=. _ERROR_=1 _N_=5
NOTE: 欠損値を含んだ計算により、以下の箇所で欠損値が生成されました。
      (回数)(行:カラム)
      5 72:32
NOTE: データセットWORK.EXAM_PERSONから5オブザベーションを読み込みました。
NOTE: データセットWORK.EXAM_RESULTは5オブザベーション、5変数です。
```

図 6. 実行フェーズでの log 出力例（数値の演算を行う処理に文字値[*1] が含まれている場合）

*1 数値の演算を行う処理に文字値（SEMESTER 変数）が含まれている場合、LOG ウィンドウで NOTE として「数値への変換もできず，演算が不可能だったために欠損値が格納された」旨を説明する。さらに、_ERROR_変数や_N_変数などの PDV 中の格納値も log で説明がなされる。error を確認した場合には、LOG ウィンドウでも_ERROR_変数に 1 が格納されることが確認できる。

このように、実行フェーズではエラーを確認の上でオブザベーションを 1 つずつ作成し、データセットを出力している。特に図 5 で意識してほしいことは初期化の条件についてである。これらを意識することが以降で説明するデータセットマージ注意事例の理解を深めることにつながる。

3　DATA ステップ処理を考慮したデータセットマージ注意事例

3.1 (見かけ上) 理論と異なる結果を得るマージ事例

前章では、PDV と DATA ステップ処理について確認を行った。それらを踏まえたうえで応用事例として、データセットマージで見かけ上理論と異なる結果を得るマージ事例を紹介する。
例えば、exam_score と exam_person の 2 つのデータセットをマージさせて、SEMESTER 変数の格納値を exam_score に格納されている”Q1” に上書き、もう 1 つが SCORE 変数について「BORDER 以上であれば合格を表す数値 1 を BORDER 変数に格納」し、「BORDER 未満であれば不合格を表す数値 0 を BORDER 変数に格納する」ためのプログラムを作成する（図 7）。

exam_person

CLASS	NAME	SEMESTER	SCORE
A	XXX	Q-1	45
A	YYY	Q-1	35
B	XXX	Q-1	60
B	YYY	Q-1	35
B	ZZZ	Q-1	50

exam_score

CLASS	SEMESTER	BORDER
A	Q1	40
B	Q1	60

pass_fail

CLASS	NAME	SEMESTER	SCORE	BORDER
A	XXX	Q1	45	1
A	YYY	Q1	35	0
B	XXX	Q1	60	1
B	YYY	Q1	55	0
B	ZZZ	Q1	50	0

図 7. マージ事例（想定している出力結果）

このような目的で例えば図 8 のようにプログラムを作成してデータセット出力を行うかもしれないはその場合は図 7 のような出力結果は得られない。

```
/*Program*/

data pass_fail;
   merge exam_person exam_score;
   by CLASS;
   if    BORDER <= SCORE  then BORDER = 1;   /*1 : PASS SCORE*/
   if . < SCORE  <  BORDER then BORDER = 0;   /*0 : FAIL SCORE*/
run;
```

pass_fail

CLASS	NAME	SEMESTER	SCORE	BORDER
A	XXX	Q1	45	1
A	YYY	Q-1	35	1
B	XXX	Q1	60	1
B	YYY	Q-1	35	1
B	ZZZ	Q-1	50	1

図 8. マージ結果（作成したプログラムと得られる出力結果 *1）

*1 出力データセットの赤枠部分は図 7 で想定していた格納値と異なる結果を得たものを示している。

図 8 のように SEMESTER 変数の格納値が上書きできない理由、そして BORDER 変数に想定と異なる結果が得られる理由を説明する。

3.2 事例発生理由

想定と異なる結果が得られる理由を説明する前に、まずは MERGE ステートメントによるデータセットマージの特徴について説明する。ここでは 2 データセットの横結合を想定して説明している。

MERGE ステートメントは BY ステートメントで指定した変数の格納値ごとにレコードをグループ分けし、特定グループのレコードに対して横結合を行っている。この横結合は「必ずしもすべてのレコードが横結合されるわけではないこと」に留意いただきたい。BY ステートメントで指定した変数で 1 対 1 のレコードで横結合が出来るのであれば、レコード同士で結合が実施できる（図 9-[b]）。しかしながら、1 対多のような横結合を行う場合には、特定グループのレコードに対して「最初に登場したレコードのみに対して」横結合を行っているのである（図 9-[c]）。

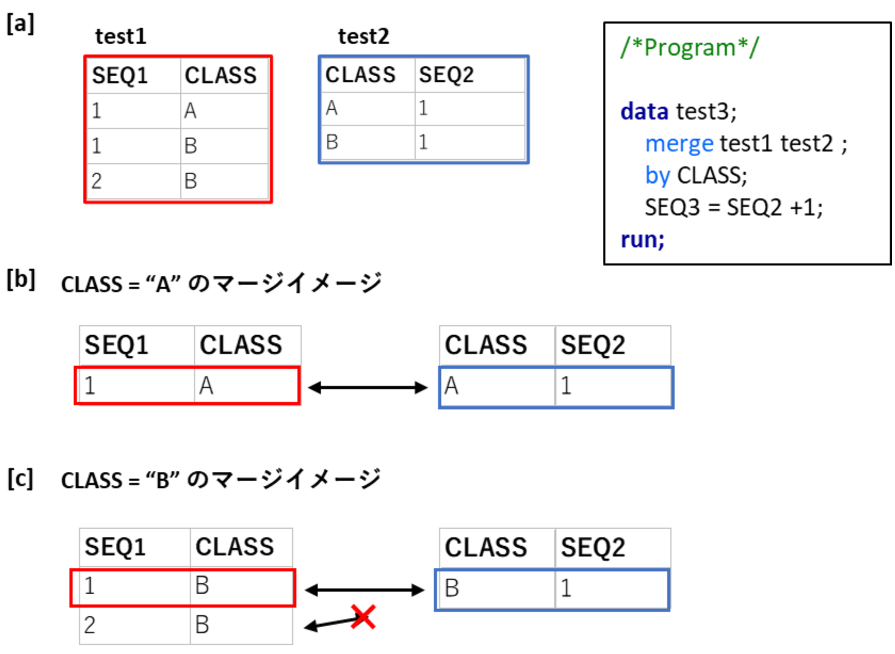


図 9. MERGE ステートメントによる横結合のイメージ *1

[a]イメージの説明に用いられるデータセットとプログラム

[b] 1 対 1 のレコードにおける横結合

[c] 1 対多のレコードにおける横結合

*1 [b], [c]のレコードにおいて色のついた枠線を付与したレコードは横結合が行われたレコードを表す。

図 9-[c]のイメージのように、図 8 でも各クラスのレコード群の 1 行目のみとレコードのマージが行われなかったため SEMESTER 変数の値が上書きされなかったのである。

続いて、BORDER 変数の格納値に関する問題が発生する理由について説明する。

この原因については、前章の図 5-[c']における格納値の初期化でも説明した「初期化の対象となる変数は、SET ステートメントで指定したデータセットのレコードではなく、プログラムで割り当てた変数」であることを踏まえて MERGE ではどのように初期化がなされているのか考えることで理由が理解できるだろう(図 10)。

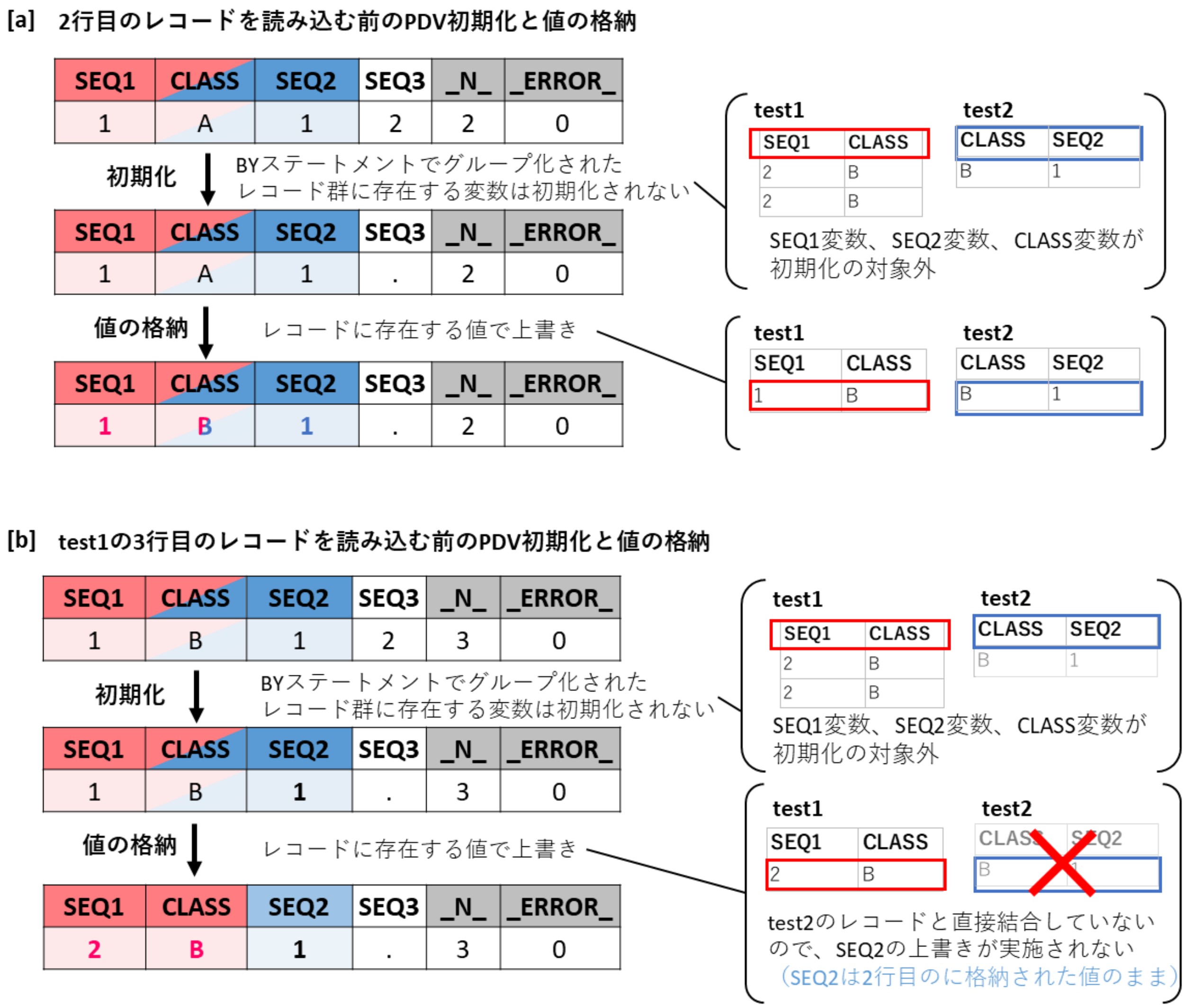


図 10. MERGE ステートメント処理における 1 対多レコードでの PDV 初期化のイメージ *1 (図 9-[a]のプログラムを使用)

[a] 1 対多の横結合での 1 行目の結合イメージ

[b] 1 対多の横結合での 2 行目以降の結合イメージ

*1 赤色の枠・塗りつぶし部分は test1、青色の枠・塗りつぶし部分は test2 のデータ・変数を表す。

MERGE ステートメントでは、2 行目以降のレコードを読み込む際には、BY ステートメントでグループ化されたレコードに存在する変数は初期化されないようになっており、初期化されなかった変数はレコードの格納値で上書きされる（図 10-[a]）。2 行目以降の処理についても、PDV 初期化の対応は同様であるが、MERGE ステートメントの仕様上、横結合が実施されておらず値が格納できない。そのため、初期化されなかった値がそのまま格納値として残ってしまう（図 10-[b]）。このように、MERGE ステートメントでの横結合はステートメントの特性と PDV の初期化の都合により、「見かけ上」横結合ができているのである。図 8 では初期化の対象外の変数に導出した値を格納したため、想定と異なる結果を得たのである。

これにより、MERGE ステートメントを扱う際の注意事項が、①各グループの 1 行目のレコード同士でしか横結合が実施されない、②変数の初期化は BY ステートメントでグループ化されたレコードに存在する変数は対象外である、ことが理解できるだろう。

図 8 のような誤りを防ぐために、複数データセット内で同一変数がないか確認して、必要に応じて RENAME する、導出した値は新規変数を用いて作成するなどして処理が明確になるように対応すべきである（図 11）。

```
/*Program*/

data pass_fail;
   merge exam_person (drop=SEMESTER)             /*Mergeによる上書きを適切に行うためにdropを使用*/
       exam_score  (rename=(BORDER=_BORDER)) /*BORDER変数に適切に値を格納するために、変数名を変更*/
   ;
   by CLASS;
   if   _BORDER <= SCORE    then BORDER = 1;     /*1 : PASS SCORE*/
   if . < SCORE    < _BORDER then BORDER = 0;     /*0 : FAIL SCORE*/
   drop _BORDER;                                /*exam_scoreでrenameした変数を削除*/
run;
```

pass_fail

CLASS	NAME	SCORE	SEMESTER	BORDER
A	XXX	45	Q1	1
A	YYY	35	Q1	0
B	XXX	60	Q1	1
B	YYY	35	Q1	0
B	ZZZ	50	Q1	0

図 11. 適切なマージを行うためのプログラム例と出力結果

なお、図 8 で実行したプログラムの PDV でのデータ処理イメージについては、以下の図 12 に示すプログラムを用いて処理した内容の可視化ができるので自身でも確認いただきたい。

```
/*Program*/

data pass_fail;

   _out = 1; output;

   merge exam_person
         exam_score
   ;
   by CLASS;

   _out = 2; output;

   if    BORDER <= SCORE  then BORDER = 1; /*1 : PASS SCORE*/
   if . < SCORE  <  BORDER then BORDER = 0; /*0 : FAIL SCORE*/

   _out = 3; output;

run;
```

図 12. PDV のデータ処理イメージ（図 8 のプログラム[*1] を使用した場合）

*1 OUTPUT ステートメントを用いて、PDV での処理イメージをデータセット出力できるようにしている。

4　まとめ

DATA ステップ処理は図 1 で説明しているように準備、実行されている。データセットの作成は 1 レコードずつ PDV の初期化や値の格納が実施されることで対応される。しかしながら、初期化の対象が特定の変数で実施されていることを理解できていないと想定しない結果を得ることにつながる。

このように、データステップ処理のような基本的な処理においても不可視の挙動が多く存在し、それが起因して思うような結果が得られない場合があることが理解できただろう。プログラミング初心者においては、不可視の部分を「なんとなく動く」という認識で対応されていることも多いが、誤った結果の出力につながりかねない。想定通りの出力結果が得られるプログラムが作成できるよう、本稿の内容だけでなく SAS 公式テキスト「Base Programming Using SAS® 9.4　完全ガイド」や SAS® Help Center を確認の上で、PDV 等の挙動だけでなく具体的にどのように値が格納されるのか理解を深めていただきたい。

また、「処理した内容の可視化」は、プログラミングにおけるトラブルシューティングの時間を短縮するうえで重要な役割を果たす。本稿が SAS 初心者の方にとって、ステートメントやプロシジャの理解を深めるためにも、得られたデータセット（場合によっては LOG ウィンドウ）でプログラムの実行内容および実行結果を確認することの重要性についても理解する一助となることを望む。

5　参考文献

・SAS Certified Specialist Prep Guide: Base Programming Using SAS 9.4, 2019, SAS Institute Inc., Cary, NC, USA

・Dalia C. Kahane (2011), “SAS® DATA Step – Compile, Execution, and the Program Data Vector” Proceedings of the 2011 North East SAS Users Group Conference

・”DATA ステップの処理の概要”, SAS® Help Center (9.4 Base SAS 言語リファレンスの新機能: 概念), <https://documentation.sas.com/doc/ja/lrcon/9.4/p08a4x7h9mkwqvn16jg3xqwfxful.htm>,（参照日：2023/08/30）

・”SET Statement”, SAS® Help Center (SAS® Visual Data Mining and Machine Learning 8.1),
<https://documentation.sas.com/doc/da/vdmmlcdc/8.1/lestmtsref/p00hxg3x8lwivcn1f0e9axziw57y.htm>,
（参照日：2023/08/30）

・” MERGE Statement”, SAS® Help Center (SAS® Visual Data Mining and Machine Learning 8.1),
<https://documentation.sas.com/doc/da/vdmmlcdc/8.1/lestmtsref/n1i8w2bwu1fn5kn1gpxj18xttbb0.htm>,
（参照日：2023/08/30）

・井上　貴博 (2017), “初/中級者が陥りやすいプログラミングエラー”, SAS ユーザー総会 2017

バッチサブミットについて

○大山暁史

（イーピーエス株式会社）

Introduction of batch submit

Akifumi Oyama
EPS Corporation

要旨

SAS のバッチ実行は有用な機能だが、使い方についてまとめられた資料が少ないと思われる。SAS ユーザーの利用の幅を広げるため、本発表ではバッチ実行の長所や短所、基礎的な使用法などを提示する。

キーワード：Batch Submit

バッチサブミットとは

バッチサブミットとは、SAS ウィンドウを起動せずに SAS プログラムを実行することである。以下に Windows 環境下でのバッチサブミットの実行方法例と、対話実行（SAS ウィンドウを起動する Display Manager System (DMS)モードでのプログラム実行）と比較したバッチサブミットのメリット・デメリットを示す。

バッチサブミットの実行方法例

バッチサブミットは主に下記の方法で実行できる[1] [2]。

方法① プログラムを右クリック→「SAS でバッチサブミット」を押下
方法② 用意したバッチファイルをダブルクリック
方法③ SAS プログラム内の SYSTASK ステートメントでの実行

バッチサブミットのメリット

バッチサブミットのメリットとして、下記を挙げる。

・work フォルダやマクロ、ライブラリが初期化された状態で実行される
・プログラム内で出力処理を記載しなくとも log, lst を出力可能
・プログラム実行中でも DMS モードでのプログラム開発を並行できる
（例えば処理に数時間かかるプログラムを DMS モードで実行すると処理が終了するまで DMS モードで作業できないが、バッチサブミットをすると数時間の処理を実行している間にも DMS モードで作業できる）

・複数のプログラムを並行して実行できる
・ログやアウトプットに色や太字フォント等を用いない分、処理時間が小さくなる
・autoexec.sas が自動実行されるので、SAS ファイルを%include する必要がない
・Windows のタスクスケジューラと組み合わせると予約実行や定時に繰り返し実行することができる

バッチサブミットのデメリット

バッチサブミットのデメリットとして、DMS 用の機能である dm コマンドや SAS_EXECFILEPATH 環境変数が使用できないことや、gplot プロシジャがうまく実行できないことなど、対話実行と挙動が変わることが挙げられる。

バッチサブミット使用時の留意事項

以下にバッチサブミットを使用する上で留意すべき事項について記載する。

autoexec.sas の自動実行

バッチサブミットする SAS プログラムと同じフォルダに autoexec.sas という名前のプログラムを用意しておくと、バッチサブミットをしたプログラム内の処理より前に autoexec.sas で記載されている処理が自動実行(automatic execution)される。

そのため、libname ステートメント等、環境設定周りの処理を autoexec.sas に記載しておくと便利である。一方で autoexec.sas に不適切な処理があるとバッチサブミットしたプログラムの処理に影響を与える可能性があるので注意が必要である。

SAS ウィンドウを閉じないと WARNING が出る

SAS ウィンドウを立ち上げていると、ユーザー情報がロックされている状態となる。バッチサブミットで実行されたプログラムでユーザー設定を変えようとしても、SAS ウィンドウを立ち上げた状態であると変更が反映されないため下記の WARNING が出る。そのためバッチサブミット時には SAS ウィンドウを閉じるべきである。

```
WARNING: SASUSER レジストリを WORK レジストリにコピーできませんでした。このセッション中のレジストリのカスタマイズはできません。
```

ods output 使用時の注意

DMS モードではデフォルトで ODS Graphics が ON となっているため、ods graphics on; の記載なしで ods output が処理されるが、バッチサブミットではデフォルトで ODS Graphics が OFF となっているため、ods graphics on; を明示しないと適切に ods output が処理されないので注意が必要である[3]。

```
ods graphics on;
ods output Survivalplot=Survivalplot_test;
proc lifetest data=test
   method=km plots=survival;
   time Time * status(0) ;
   strata Gender;
run;
```

ODS RTF の空白行削除

ODS RTF でデータセットを rtf 出力する処理があり、出力基データセットを data _NULL_ ; で作成している場合、バッチサブミットを行うと rtf の空白行が削除される場合がある。

DMSモードでの実行			バッチサブミット		
Gender	Number in Class	Percent of total (%)	Gender	Number in Class	Percent of total (%)
女子	9	47.37	女子	9	47.37
男子	10	52.63	男子	10	52.63

バッチサブミットを行う際に空白行を削除したくない箇所には下記の処理を追記すると空白行削除を回避できる[4]。

```
put '(*ESC*)¥~';
```

バッチサブミット周りのテクニック

バッチサブミットを効率的に実施する際の手法について紹介する。

バッチファイルの作成方法

複数ファイルを一括でバッチサブミットしたり、log の出力場所を制御したりする場合などは前述のバッチサブミットの実行方法例の方法②が便利である。ここでは windows バッチファイル（bat ファイル）の作成方法について述べる。

＜作成手順＞

① 適当なテキストファイルを用意し、拡張子を.bat に変更する

② 下記記載例に倣い、必要に応じて各オプションを指定する[5]

＜バッチファイル記載例＞

```
"[フォルダパス]¥sas.exe" -sysin "[フォルダパス]¥[ファイル名].sas " -config "[フォルダパス]¥sasv9.cfg"
```

1 つのプログラムについてのオプションを 1 行で記載する。複数プログラムを一括で実行したい場合には、実行するプログラムの分だけ行を増やす。

＜各オプションについて＞

-sysin オプション

実行対象の SAS プログラムのパスとファイル名を指定するオプション。

-config オプション

SAS の環境設定ファイル(sasv9.cfg)を指定するオプション。複数の SAS が存在する環境の場合、指定する環境設定ファイルが目的の SAS のものであるか確認する必要がある。なお、環境設定ファイルの格納場所の調べ方については下記の通りである。

Windows を SAS で検索
→ ファイルの場所を開く
→ 開いたフォルダの SAS ショートカットを右クリックでプロパティ
→ リンク先欄

-log オプション

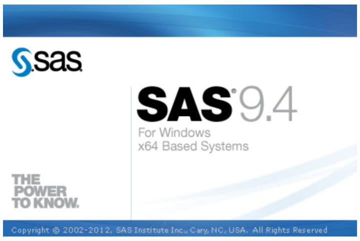
ログファイルの出力先とファイル名を指定するオプション。ファイル名を指定しない場合、ファイル名は「プログラム名.log」になる。また、オプションを指定しない場合、ログファイルはプログラムファイルが保存されているフォルダに出力される。

-print オプション

ファイル名を指定しない場合、ファイル名は「プログラム名.lst」になる。また、オプションを指定しない場合、アウトプットファイルはプログラムファイルが保存されているフォルダに出力される。

-nologo オプション

バッチファイル実行時に SAS のロゴを出さないようにするオプション。



-nosyntaxcheck オプション

構文エラーがあっても構文チェックモードを有効化させないようにするオプション。後続の処理が止まってしまうことを回避できる。

-sysparm オプション[6]

　バッチプログラム内に記述したパラメータを SAS プログラム内の処理に受け渡すことができるオプション。-sysparm オプションの引数には最大 200 バイトまでの文字列を指定できる。

　例えばバッチファイルに -sysparm JPNSAS と記載して実行すると SAS プログラムが実行される際に SAS プログラム内の&sysparm に JPNSAS を展開してくれる。

–autoexec オプション

　バッチファイルのコマンドにて –autoexec [プログラム名].sas を入力すると autoexec.sas の代わりに別プログラムを自動実行してくれる[7]。

下記の例だと、 prg1.sas の実行前に before.sas が自動実行される。

※スペースの関係上複数行に分けて記載しているが、バッチファイルには 1 行で記載する。

```
"[フォルダパス]¥sas.exe"
-autoexec [フォルダパス]¥before.sas
-sysin "[フォルダパス]¥prg1.sas"
-config "[フォルダパス]¥sasv9.cfg"
```

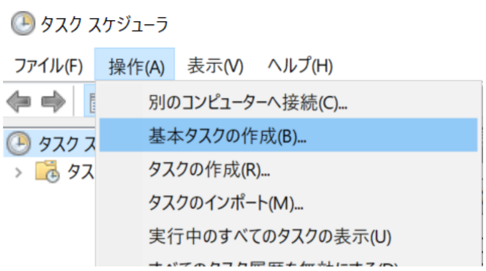
予約実行・繰り返し実行

　バッチファイルを用意し、Windows タスクスケジューラで予約実行するように設定すると任意の時間にバッチサブミットを行うことができる。

＜Windows タスクスケジューラの設定方法＞

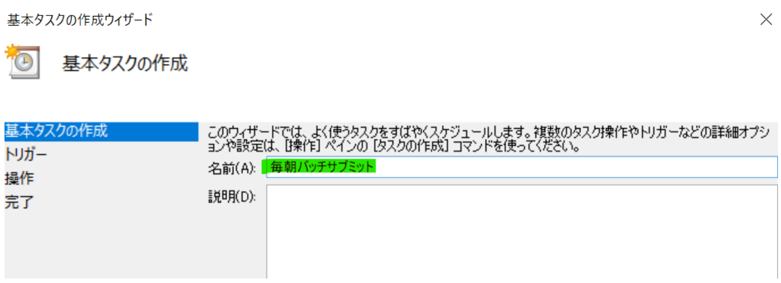
Windows 内を「タスク スケジューラ」で検索

→「操作」タブの「基本タスクの作成」を選択

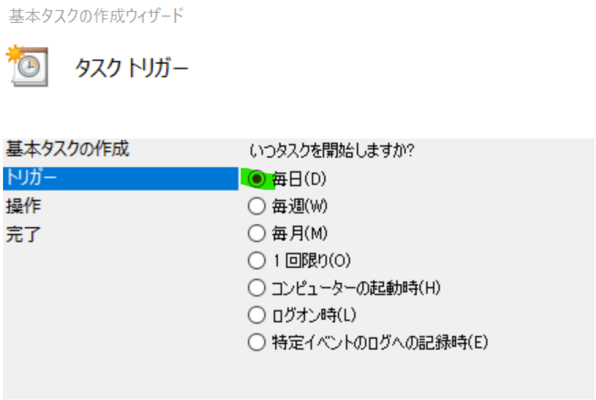


→「基本タスクの作成ウィザード」を入力する

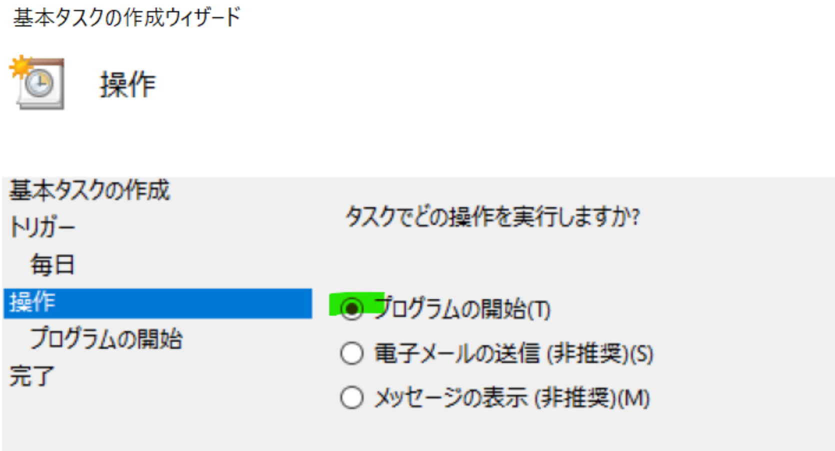
　　「基本タスクの作成」: 任意のタスク名を入力する

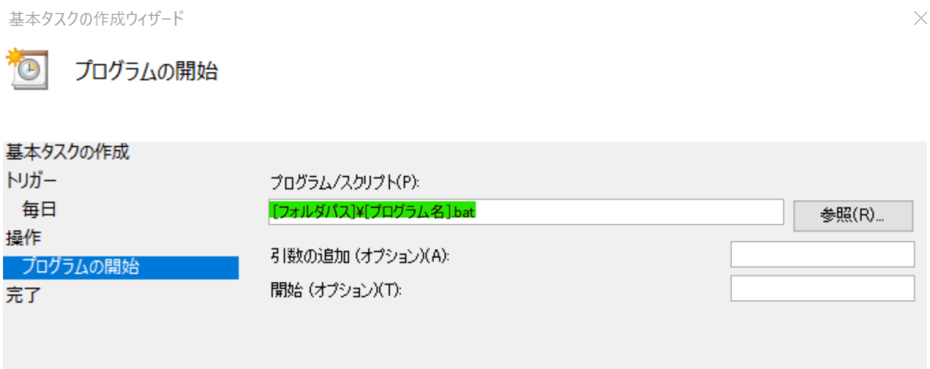


「トリガー」: バッチサブミットしたい任意のタイミングを入力する



「操作」:「プログラムの開始」を選択し、実行するバッチファイルを選択する





なお、タスクスケジューラでバッチファイルを実行する際には、バッチファイル内で-autoexec オプションによる autoexec.sas の指定を行わないと autoexec.sas を読み込まないようなので注意が必要である。

```
-autoexec [フォルダパス]¥autoexec.sas
```

SYSTASK ステートメントによるバッチサブミット

SAS プログラム内に SYSTASK ステートメントを用いた下記のような記載をすると、SAS プログラム実行時にバッチサブミットを行うことができる[2][8]。

＜プログラム記載例＞

```
SYSTASK command
" '[フォルダパス]¥sas.exe' -sysin '[フォルダパス]¥[ファイル名].sas ' -config '[フォルダパス]¥sasv9.cfg' "   wait;
```

SYSTASK command でのバッチサブミットを応用するとタイムアウト処理やリターンコード取得による処理分岐を行うことができる。

＜タイムアウト処理＞

バッチ実行が指定の時間を超えた際に SYSTASK KILL ステートメントによってバッチ実行を強制終了することができる。下記に記載例を示す。

```
SYSTASK command " '[フォルダパス]¥sas.exe' -sysin '[フォルダパス]¥[ファイル名].sas ' -config '[フォルダパス]¥sasv9.cfg' " mname=Taskname ;

WAITFOR &Taskname TIMEOUT=120; /* Number of seconds to wait before assuming timeout */
%if &SYSRC ne 0 %then %do; /* Task timed out */
  %put "Batch SAS step timed out";
SYSTASK KILL &Taskname;
%end;
```

＜リターンコード取得処理＞

SYSTASK ステートメントによるバッチサブミットを行うと、バッチサブミットされたプログラムの実行結果に対応してリターンコードを取得することができる。

表 プログラム実行結果と対応するリターンコード

Condition	Return Code
All steps terminated normally	0
SAS System issued warning(s)	1
SAS issued error(s)	2
ABORT statement	3
ABORT RETURN statement	4
ABORT ABEND statement	5

下記のような記載でバッチサブミットを行うと、実行内容に問題があった場合には任意のアラートをログに出力させることができ、不備を検知できる。

```
SYSTASK command " '[フォルダパス]¥sas.exe' -sysin '[フォルダパス]¥[ファイル名].sas ' -config '[フォルダパス]¥sasv9.cfg' " status=Taskrc wait;

%if &Taskrc > 1 %then %put [任意のアラート文];
```

また、取得したリターンコードが異常を示す値であった場合、下記のようなコードを記載することで任意のメールアドレスにアラートメールを送付することも可能である。シミュレーションプログラム等、実行時間の長いプログラムを実行する際などに SYSTASK command を用いたバッチサブミットが有用であると考えられる。

```
SYSTASK command " '[フォルダパス]¥sas.exe' -sysin '[フォルダパス]¥[ファイル名].sas ' -config '[フォルダパス]¥sasv9.cfg' " status=Taskrc wait;

%if &Taskrc > 1 %then %do;
  filename notify email "[送信先のメールアドレス]"
  subject="[メールタイトル]" ;
  data _null_;
    file notify;
    put "[メール本文]" ;
  run;
%end;
```

終わりに

本発表ではバッチサブミットの長所や短所、基本的な使用法についてまとめた。バッチサブミットと対話実行のそれぞれの特徴を理解した上で実行方法を選択することが重要であると考える。本発表が SAS ユーザーの研究・業務の効率化につながる議論の基となれば幸いである。

引用文献

[1] "SAS をバッチモードで実行する方法"
https://wikiwiki.jp/cattail/SAS%E3%82%92%E3%83%90%E3%83%83%E3%83%81%E3%83%A2%E3%83%BC%E3%83%89%E3%81%A7%E5%AE%9F%E8%A1%8C%E3%81%99%E3%82%8B%E6%96%B9%E6%B3%95
(Accessed Aug 12, 2023)

[2] Denis Cogswell, "More Than Batch – A Production SAS® Framework", SUGI 30 Applications Development - Paper 021-30 021-30: More Than Batch: A Production SAS® Framework (Accessed Aug 21, 2023)

[3] Aakar Shah, “A Beginner’s Guide to Create Series Plots Using SGPLOT Procedure: From Basic to Amazing”, PharmaSUG 2022 - Paper QT-169 https://www.lexjansen.com/pharmasug/2022/QT/PharmaSUG-2022-QT-169.pdf (Accessed Aug 12, 2023)

[4] Irina Walsh, “Pros and Cons of Interactive SAS® Mode vs. Batch Mode” https://www.lexjansen.com/wuss/2010/coders/2937_4_0_COD_Walsh.pdf (Accessed Aug 12, 2023)

[5] SAS FAQ “Windows 版 バッチ（非対話）モードでの実行” https://www.sas.com/offices/asiapacific/japan/service/technical/faq/list/body/pc068.html (Accessed Aug 12, 2023)

[6] Technical Support “バッチ処理の際にパラメータを渡したい” https://www.sas.com/offices/asiapacific/japan/service/technical/faq/list/body/ba124.html (Accessed Aug 12, 2023)

[7] SAS 9.2 Documentation “Files Used by SAS” https://support.sas.com/documentation/cdl/en/hostwin/63285/HTML/default/viewer.htm#a000104286.htm#autoexec (Accessed Aug 12, 2023)

[8] SAS 9.2 Documentation “SYSTASK Statement: Windows” https://support.sas.com/documentation/cdl/en/hostwin/63285/HTML/default/viewer.htm#win-stmt-systask.htm (Accessed Aug 21, 2023)

G-formulaによるtime-varying treatmentsの因果効果の推定

○鈴木 徳太[1], 岡本 憲曉[2], 折原 隼一郎[3]

（[1]東京医科大学大学院医学研究科, [2]慶應義塾大学大学院経済学研究科,

[3]東京医科大学医療データサイエンス分野）

Estimation of causal effects of time-varying treatments using the g-formula

Norihiro Suzuki[1], Noriaki Okamoto[2], Shunichiro Orihara[3]

[1]Graduate school of Medicine / Tokyo Medical University

[2]Graduate school of Economics / Keio University

[3]Department of Health Data Science / Tokyo Medical University

要旨

時間依存性治療の因果効果を推定する場合，一般に時間依存性交絡の問題が生じる．時間依存性交絡を適切に調整し興味のある因果効果をバイアスなく推定する手法として g-methods と総称される 3 つの手法が提案されている．本稿では, g-methods の中でも特に用いられることの多い周辺構造モデルにおける IPTW 法と g-formula の基本的コンセプトを簡潔に説明する．

キーワード：統計的因果推論，時間依存性治療，時間依存性交絡, g-formula, IPTW of MSMs, g-methods

1. はじめに

統計的因果推論とは「原因」と「結果」の学問である．すなわち，ある治療のような特定の原因が興味のある結果（アウトカム）に与える因果効果を定量的に評価することが最大の目的となる．近年の統計的因果推論は Judea Pearl により提案された構造的因果モデル [1] と, Jerzy Neyman により提案され Donald B. Rubin によりその後体系化された Neyman-Rubin 因果モデル [2, 3] という 2 つの枠組みに立脚し議論がされる．前者は変数間の関係（生成過程）を，構造方程式を用いて表現するものであり，特に変数の関係を，構造方程式をもとにして視覚的に示した因果ダイアグラムが議論に用いられている．後者は潜在アウトカムモデルや反事実アウトカムモデルとも呼ばれ，これは「もし〇〇という治療が行われたら」という仮想的な状況において定義されるアウトカム（潜在アウトカム）を導入することによって，関心のある因果効果を定義し，諸仮定の下でデータからそれを識別する．統計的因果推論の理論は従来から医学，経済学分野を中心として発展し

てきたが，最近では関連する書籍が多く出版されることに代表されるように言葉自体が市民権を得てきているといっても良いだろう．

入門的な書籍や文献において，興味のある治療がベースラインでのみ行われる設定を扱うことが多い．しかし，実際に扱う問題では，ベースライン以降の複数時点に渡って逐次的に治療が行われる状況を想定することが多いと考えられる．そのため逐次的に行われる治療（sequential treatments）に対する因果推論の理論は数学・統計学的に比較的高度となるものの，アプローチを理解することには一定の価値がある．厳密な定義については後述するが，逐次的に行われる治療はそのレベルが時間に依存せずに定まるか，それとも時間に依存して定まるかによって二分化される.[4,5] 具体的には，前者は単一時点でのみ行われる治療と併せて時間固定性治療（time-fixed treatments）と，後者は時間依存性治療（time-varying treatments）と呼ばれる．この分類からも類推されるように，治療が複数回行われるとしても治療レベルが時間に依存しない場合には，ベースラインでのみ治療が行われる場合と同様の手法が因果効果の推定に適用可能である．しかし，時間依存性治療の因果効果の推定を行う場合，一般に時間依存性交絡（time-dependent confounding）と呼ばれる問題が発生し，層別化や回帰，マッチングといった，治療がベースラインでのみ行われる状況で標準的に利用される手法を用いると推定結果にバイアスが含まれる．時間依存性交絡に適切に対応した上で時間依存性治療の因果効果を推定する手法としては, James M. Robins によって体系化された generalized methods（以降, g-methods と呼ぶ）と総称される(i) g-computation algorithm formula (g-formula), (ii) 周辺構造モデルにおける IPTW 法, (iii) 構造ネストモデルにおける g-estimation という 3 つの手法が基本となる.[4,5] 時間依存性治療，及び g-methods に関する日本語での資料は非常に少ないものの，近年では日本製薬工業協会医薬品評価委員会データサイエンス部会が解説資料を提供している.[6]

本論文では上記の資料を踏まえた上で時間依存性治療や時間依存性交絡，および(i)と(ii)の統計的理論を簡潔に説明する．また g-methods の中でも特に用いられることの両手法は個別で議論される場面が多いため，SAS での実装方法に加えてシミュレーションベースでの比較結果を SAS ユーザー総会 2023 において紹介する．なお本稿では興味のある原因は特定の治療とするが，曝露や介入といった用語と同義であることに注意されたい．

2. 治療と治療レジメン

2.1 本稿を通じた仮定

ベースライン時点，及びそれ以降の時点での治療による因果効果の推定にのみ関心があるものとし, N 人で構成される研究対象集団は十分に定義された閉鎖コホートであるとする．また，各被験者に関しては共変量，治療が一定間隔（e.g., 隔週）で測定され，観察打ち切りや脱落，測定誤差は存在しないとする．さらに興味のあるアウトカム（連続値 or 二値）は最終時点でのみ測定される単一の変数であるとする．

2.2 時間固定性治療（Time-fixed treatment）

時間依存性治療に対する因果推論の議論を行うにあたり，まずはある治療が時間に依存しない時間固定性治療（time-fixed treatments）の状況を議論する．ある治療が時間に依存しない, time-fixed であるとは，研究対象集団に含まれるすべての被験者に関して各々のベースライン治療のレベルがその後すべての時点における治療レベルを決定することを意味する．この状況としては次の 3 つが想定される.[4]

1. ベースラインでのみ治療が行われる
2. ベースラインでの治療レベルが時間経過によって不変
3. 決定論的に各時点の治療レベルが定まる

ここで，Aをベースラインでの二値の治療変数（1: treated, 0: untreaded），Yをアウトカム，Lをベースライン共変量（一般にはベクトル）とする．なお，アルファベットの大文字は確率変数を，小文字はその実数値であることを意味する．さらに，仮想的にある介入（$A = a$）を受けた場合に観測されるアウトカム（潜在アウトカム）を$Y^a \in (Y^{a=1}, Y^{a=0})$と表現する．これを用いて加法的なスケールでの個別因果効果，並びに平均因果効果は期待値記号Eを用いて以下のように定義される．

個別因果効果（individual causal effect）

$$Y^{a=1} - Y^{a=0}$$

平均因果効果（average causal effect）

$$E[Y^{a=1} - Y^{a=0}]$$

潜在アウトカムは，各個人でいずれか一つしか現実には観測がされないため，個別因果効果は識別することができない．つまり，試験治療を受けた被験者は$Y^{a=1}$のみ，対照治療を受けた被験者は$Y^{a=0}$のみ観測される．これは潜在アウトカムを用いた因果推論における根源問題（fundamental problem）[7]として知られており，このため一般には平均因果効果が推定対象となる．

上記の 2, 3 の状況，治療時点が複数回あるとしてもそのレベルが時間に依存しない場合には，平均因果効果の定義に治療の行われた時間を参照する必要はない．これはすべての被験者に対し，ベースライン時点の治療によってその後の治療レベルが決定されるためである．つまり，時間固定性治療の因果効果を推定するにあたっては 1~３の状況を区別する必要はなく，統計学的には同じ扱いをすることができる．詳細については参考文献を参照されたい.[4] 平均因果効果を推定するためには次の 3 つの識別条件（identifiable assumptions）が必要である．

識別条件（identifiable assumptions）

- 一致性（consistency）

$$\text{If } A = a \text{ for a given subject, then } Y^a = Y \text{ for that subject}$$

特にAが二値である時

$$Y = AY^{a=1} + (1 - A)Y^{a=0}$$

- 条件付き交換可能性（conditional exchangeability）

$$Y^a \coprod A|L = l \text{ for each possible value } a \text{ of } A \text{ and } l \text{ of } L$$

$\coprod$は統計学的な独立を意味し，未測定交絡が存在しないこと（no unmeasured confounding）と同義である．

- 正値性（positivity）

$$\text{If } f_L[L = l] \neq 0, \text{ then } f_{A|L}[a|L = l] > 0 \text{ for all } a$$

$f_L[L=l]$はLの周辺確率密度関数，$f_{A|L}[a|L=l]$はLが与えられた下でのAの条件付き確率密度関数を意味する．なお確率密度関数は適宜確率関数として表現可能である．

上記の識別条件は完璧に遵守された理想的なランダム化比較試験においては，試験デザイン上その成立が認められるが，観察研究においては識別条件をあくまで“仮定する”こととなり，その成立を認めることが妥当であるか議論が必要となる．識別条件がすべて成立する場合には時間固定性治療に対しては後述する時間依存性交絡の問題は発生せず，ベースライン共変量の調整に基づく手法によって平均因果効果を推定することが可能である．ただし交絡調整に必要となる共変量が正しく特定・測定され，調整に統計モデルを用いる場合にはモデルの誤特定がないことが必要となることに注意されたい．時間固定性治療に対する手法の解説，及び上記の仮定の解釈については多くの文献で扱われているため本稿では取り扱わない．

2.3　時間依存性治療（Time-varying treatment）/ 治療レジメン（treatment regime）

時間依存性治療とは時間固定性治療以外の治療であり，治療が行われる時点ごとに異なるレベルを取りうる治療を意味する．ここで先ほどの記法に加え$t=0,\dots,K$を時点，A_tを時点tにおける治療，L_tを時点tにおける共変量，そしてオーバーバー（￣）を用いてそれらの歴を表すものとする．すなわちA_0, L_0はそれぞれベースラインでの治療と共変量であり，$\bar{A}_t=\{A_0,A_1,\dots,A_t\}$，$\bar{L}_t=\{L_0,L_1,\dots,L_t\}$は時点$t$までの治療歴，及び共変量歴を指すものとする．なお以降では単に$\bar{A}_K$，$\bar{L}_K$を単に$\bar{A}$，$\bar{L}$とも表記する．

ベースラインから時点Kまでの一連の治療を示す$\bar{A}$は，治療レジメン（treatment regime）と呼ばれ，分野や文脈によっては regime の代わりに strategy, plan, policy, protocol とも記載される．任意の治療レジメン$\bar{A}=\bar{a}$に対しても治療が時間に依存しない場合と同様に潜在アウトカムを$Y^{\bar{a}}$として考えることが可能であり，例えば$K=2$である時，想定される潜在アウトカムは以下の 4 つである．

- $Y^{(a_0=0,a_1=0)}$：$t=0,1$でともに治療を受けない場合の潜在アウトカム
- $Y^{(a_0=1,a_1=0)}$：$t=0$では治療を受け，$t=1$では治療を受けない場合の潜在アウトカム
- $Y^{(a_0=0,a_1=1)}$：$t=1$では治療を受け，$t=0$では治療を受ける場合の潜在アウトカム
- $Y^{(a_0=1,a_1=1)}$：$t=0,1$でともに治療を受ける場合の潜在アウトカム

想定される治療レジメン，及び対応する潜在アウトカムは，時点がK個存在する二値治療に対しては2^K個存在する．時間依存性治療に対する因果推論が難しくなるポイントの一つは，この平均因果効果を定義する潜在アウトカムの数が膨大になる点である．つまり，潜在アウトカムの比較によって定義される因果効果は時間固定性治療に対しては，2.2 節で導入した平均因果効果ように明らかに 1 つの形に定まるものの，時間依存性治療A_tに対してはその因果効果に興味のある 2 つの任意の治療レジメン$\bar{a}$，$\bar{a}'(\neq \bar{a})$を定める必要がある．例えば上記の例において，$t=0,1$でともに治療を受けない場合（$a_0=0$，$a_1=0$）と比べて，ともに治療を受ける場合（$a_0=1$，$a_1=1$）には平均的にどの程度因果効果があるのかに興味があるとしたときには，平均因果効果は$E[Y^{a_0=1,a_1=1}-Y^{a_0=0,a_1=0}]$として定義される．また平均因果効果の定義は$2^K$から 2 つのレジメンを選択する組み合わせの数だけ存在するが，このうち少なくとも 2 つの治療レジメン$\bar{a}$，$\bar{a}'$に関して$E[Y^{\bar{a}}-Y^{\bar{a}'}]\neq 0$となる場合に，時間依存性治療$A_t$はアウトカム$Y$に平均的に因果効果をもつという．ここまで治療レジメンは$\bar{a}$として表記を行ったが，多くの文献において，より一般的な状況でレジメンを扱う場合にはgとして表記される．本稿もその流れに習い，以降ではgと記載する場合もある．

2.4　治療レジメンの分類

時間依存性治療に対する治療レジメンgはその値の定まり方によって，決定論的（deterministic）であるか確率的（random）であるか，さらに動的（dynamic）であるか静的（static）であるかに分類される．すなわち治療レジメンは2×2の 4 つに分類可能である.[4, 5] 本節では Young et al. (2014)[7] に準拠して定義を紹介する.

まず，ある治療レジメンgが決定論的であるとは時点tにおける治療確率がそれ以前の治療・共変量歴で条件付けした場合にすべての個人に対し, 0 もしくは 1 となることを指す．つまり，$f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left[a_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_t=\bar{l}_t\right]=0,1$である．一方ですべての決定論的ではない治療レジメンは確率的であり，確率的な治療レジメンに対しては，$0<f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left[a_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_t=\bar{l}_t\right]<1$となる．次にある決定論的治療レジメン$g$が静的であるとは，すべての時点$t$に関して$a_t$が観察される共変量歴$\bar{l}_t$のいずれにも依存せず，治療歴$\bar{a}_{t-1}$にのみ依存することを指す．すなわち，$f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left[a_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_t=\bar{l}_t\right]\equiv f_{A_t|\bar{A}_{t-1}}\left[a_t|\bar{A}_{t-1}=\bar{a}_{t-1}\right]$ となる．一方，静的でない治療レジメンは動的であるとされ，これは治療歴$\bar{a}_{t-1}$に加えて共変量歴$\bar{l}_t$によってa_tが定まる．なお動的な治療レジメンについては関数$\{g_t\left[\bar{a}_{t-1},\bar{l}_t\right];\ t=0,\dots,K\}$を用いて，$g=g_0\left[\bar{a}_{-1},\bar{l}_1\right],\dots,g_K\left[\bar{a}_{K-1},\bar{l}_K\right]$とも表される．分類の定義については一読しただけでは理解しにくいため, Young らによって紹介されている各レジメンの例を一部改変して紹介する.[7] 以下の例では研究開始時点からの日数が時点tであり, BMI が共変量L, 運動時間がAである.

- Deterministic static regime の例
  - すべての被験者に対し，各日tの運動時間を 30 分に設定する.
- Deterministic dynamic regime の例
  - t日目の開始時点の被験者の BMI が 25 以上である場合にはその日の運動時間を 30 分に設定し，そうでなければ 60 分に設定する.
- Random static regime の例
  - t日目の被験者の運動を 0.8 の確率で 30 分, 0.2 の確率で 60 分と設定する.
- Random dynamic regime の例
  - t日目の開始時点の被験者の BMI が 25 以上である場合, 0.8 の確率で 30 分, 0.2 の確率で 60 分と設定する．そうでない場合には，その日の運動時間を 60 分と設定する.

2.5　時間依存性治療に対する識別条件の拡張

時間依存性治療，及び治療レジメンの因果効果を後述する g-methods を用いて推定する際に必要となる識別条件[4, 5]を以下に示す．治療レジメンが静的であるか，動的であるかによって細部が微妙に異なるものの，各条件が意図する内容は点治療の場合と変わらない.

静的治療レジメンに対する識別条件

- 一致性（consistency）

$$\text{If }\ \bar{A}=\bar{a}\ \text{ for a given subject, then }\ Y^{\bar{a}}=Y\ \text{ for that subject}$$

- 条件付き逐次交換可能性（sequential conditional exchangeability）

$$Y^{\bar{a}} \coprod A_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_t=\bar{l}_t\ \text{ for all regime }\ \bar{a}$$

- 正値性（positivity）

$$\text{If }\ f_{\bar{A}_{t-1},L_t}\left[\bar{a}_{t-1},\bar{l}_t\right]\neq 0,\text{ then }\ f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left[a_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_t=\bar{l}_t\right]>0\ \text{ for all }\ a_t,\ l_t$$

動的治療レジメンに対する識別条件

- 一致性（consistency）

For any regime g, if $A_t = g_t(\overline{A}_{t-1}, \overline{L}_t)$ at each time t, then $Y^g = Y$ and $\overline{L}_t^g = \overline{L}_t$ for a given subject, where $\overline{L}_t^g$ is the counterfactural L-history through time t under regime g.

- 条件付き逐次交換可能性（sequential conditional exchangeability）

$$Y^g \coprod A_t | \bar{A}_{t-1} = g_t(\bar{a}_{t-2}, \bar{l}_{t-1}), \bar{L}_t = \bar{l}_t \text{ for all regime } g \text{ and } t = 0, \dots, K$$

- 正値性（positivity）

$$\text{If } f_{\bar{A}_{t-1}, L_t}[\bar{a}_{t-1}, \bar{l}_t] \neq 0, \text{ then } f_{A_t|\bar{A}_{t-1}, \bar{L}_t}[a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t] > 0 \text{ for all } (a_t, l_t)$$

3. 時間依存性交絡（time-dependent confounding）

時間依存性治療, 及び治療レジメンの因果効果を推定する場合に, 一般に時間依存性交絡（time-dependent confounding）の問題が発生する. 時間依存性交絡について紹介する前に, 点治療の状況における因果推論の前提を説明する. まずは因果 DAG と呼ばれる以下の図を参照されたい. 図中の矢線は, 研究対象集団の少なくとも 1 人に対して変数間に因果関係があること（矢線の元: 原因, 矢線の先：結果）を指す.

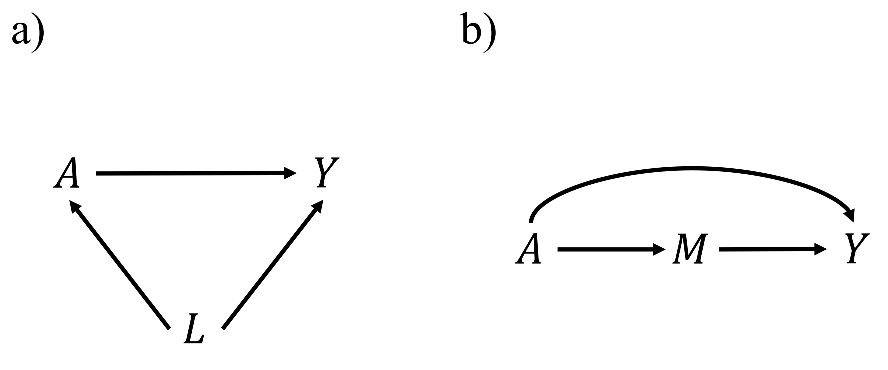


Figure 1: 交絡因子と中間因子が存在する状況を示した DAG の例

ここで Figure 1a)の状況においては, AとYの共通の原因（common cause）である変数Lが存在しており, これは交絡因子と呼ばれる. AからYの因果効果（$A \rightarrow Y$）を推定する際には, バックドアパスと呼ばれる交絡因子Lを介したAからYまでの経路（$A \leftarrow L \rightarrow Y$）があり, Lを解析に含めなければ推定結果にバイアスが含まれてしまう. また Figure 1b)の状況においては, AとYの間にMという変数が存在しており, このMは中間因子（mediator）と呼ばれる. 中間因子Mについては先ほどの交絡因子とは異なり, AからYの因果効果（$A \rightarrow Y$）を推定する際に解析に含めるとMを介した効果（$A \rightarrow M \rightarrow Y$）が除かれ, 推定結果にバイアスが含まれてしまう.

時間依存性交絡の問題とは，時間依存性治療A_tに関して交絡因子とも中間因子ともなる変数がL_t存在することによって因果効果の推定値にバイアスが含まれてしまうことである．この変数L_tは時間依存性交絡因子（time-dependent confounders, time-varying confounders）と呼ばれ，次の 2 つの条件を満たす共変量を指す.[4]

1. 1 つ前の時点の治療A_{t-1}によって引き起こされる，もしくはA_{t-1}と共通の原因をもつ
2. 次の時点の治療A_{t+1}とアウトカムYの交絡因子である（A_{t+1}に対し因果効果をもつ）

具体的に二時点で治療が行われ，時間依存性交絡因子が存在する状況の例が次の Figure 2 である．

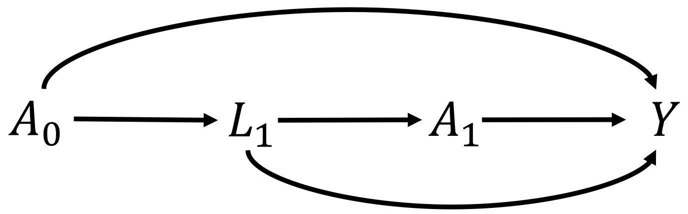


Figure 2: 時間依存性交絡因子が存在する状況を示した DAG の例

Figure 2 から明らかなように，L_1はA_0とYに関しては中間因子であり，かつA_1とYに関しては交絡因子である．すなわち，L_1を解析に含める（条件付ける）と治療レジメンの因果効果の推定結果にはバイアスが含まれることとなり，回帰やマッチング，層別化といった条件付けに基づく手法が不適切となる．先述したある時間依存性共変量が時間依存性交絡因子となる条件の 2 つ目が存在しないという仮定（no feedback と呼ばれる）の下では，回帰やマッチング，層別化を用いても一致推定量を得ることは可能である．しかしながら，大部分の医学研究においてはある時点の共変量のレベルによって次の時点の治療のレベルが決定されるため，現実的な仮定であるとは言い難い．そこで用いられるのが次節の g-methods である．

4. G-methods

冒頭で紹介したように g-methods は, (i) g-computation algorithm formula (g-formula), (ii) 周辺構造モデルにおける IPTW 法, (iii) 構造ネストモデルにおける g-estimation という 3 つの手法の総称である．本章では特に (i) と (ii) について取り上げ, (iii)については本稿の目的を逸脱するため扱わない.

4.1 G-formula

本節では決定論的動的治療レジメンに対する g-formula について簡潔に説明を行う．g-formula，並びに次節で紹介する周辺構造モデルにおける IPTW 法については，他の資料[4-6]でも扱われているため適宜そちらも参照されたい．前節で与えた識別条件がすべての決定論的動的治療レジメンgにおいて成立するとき，特に治療歴及び共変量歴（$\bar{A}_{t-1}, \bar{L}_t$）を条件付けた下で Y^g に関して条件付き交換可能性が成立している下で, $E[Y^g]$ は次の g-formula によって識別される[9]

$$\sum_{\bar{l}} E\left[Y \mid \bar{A}_K = \bar{a}_K^g = \bar{g}_K\left(\bar{a}_{K-1}^g, \bar{L}_t\right), \bar{L}_K = \bar{l}_K\right] \prod_{t=0}^{K} f\left\{l_t \middle| \bar{a}_{t-1}^g, \bar{l}_{t-1}\right\}$$

ここで総和 $\sum_{\bar{l}}$ は観測されうるすべての共変量歴について取っており，$\sum l_K \dots \sum_{l_1} \sum_{l_0}$ と書き下すことができる．治療・共変量が離散値であればアウトカムの条件付き期待値 $E\left[Y \mid \bar{A}_K = \bar{a}_K^g = \bar{g}_K\left(\bar{a}_{K-1}^g, \bar{L}_t\right), \bar{L}_K = \bar{l}_K\right]$ は, その標本平均: $n^{-1} \sum_{i=1}^{n} \hat{E}\left[Y \mid \bar{A}_K = \bar{a}_K^g = \bar{g}_K\left(\bar{a}_{K-1}^g, \bar{L}_t\right), \bar{L}_{Ki}\right]$ に置換することができ，加えてt時点での離散的

な時間依存性共変量の条件付き分布 $f\{l_t|\bar{a}^g_{t-1},\bar{l}_{t-1}\}=P\{l_t|\bar{a}^g_{t-1},\bar{l}_{t-1}\}$ をデータから推定された$\hat{P}\{l_t|\bar{a}^g_{t-1},\bar{l}_{t-1}\}$で置換することで，$E[Y^g]$をノンパラメトリックに識別可能である（non parametric g-formula と呼ぶ）．

しかし共変量が連続である場合や，離散であったとしても調整すべき共変量が複数かつ，長期間にわたり観測するケースも実際には存在する．この場合に non parametric g-formula を適用することは困難であり，条件付き期待値や条件付き分布に対してパラメトリックモデルを設定することが一般的な対応となる．例えば，追跡終了時のアウトカムの条件付き期待値については線形回帰モデルを設定し，t時点での時間依存性共変量の条件付き分布についてはロジスティック回帰モデルを当てはめることが考えられる．このようなパラメトリックモデルからの推定値を g-formula に挿入する方法が parametric g-formula と呼ばれる．

parametric g-formula は推定精度の面に関して他の手法と比較して優位性があるものの, g-null paradox という問題がある．これは sharp causal null hypothesis と呼ばれる研究対象集団全員に対して治療の因果効果が存在しないという状況において，仮に識別条件がすべて満たされている場合であってもその帰無仮説を誤って棄却してしまうという現象である.[10] g-null paradox の詳細については参考文献を参照されたい.

4.2　周辺構造モデルにおける IPTW 法

次に，決定論的動的治療レジメンにおける IPTW 法（inverse probability of treatment weighting; 治療への逆確率重みづけ法）について簡単に説明する．初めにモデルを仮定せずにノンパラメトリックに識別する方法について述べ，その後$E[Y^g]$に対してパラメトリックモデルを仮定する方法について触れる．

前節では識別条件が成立する場合に $E[Y^g]$を g-formula によって識別したが，次の式を用いても同様に識別可能であり，この識別式に基づく推定量を IPTW 推定量と呼ぶ.

$$E\left[YI\left(\bar{A}_K=\bar{a}^g_K\right)W^{\bar{A}}\right]=E\left[YI\left(\bar{A}_K=\bar{a}^g_K\right)\prod_{t=0}^{K}\frac{1}{f(A_t|\bar{A}_{t-1},\bar{L}_t)}\right]$$

ここで$I\left(\bar{A}_K=\bar{a}^g_K\right)$ はある個人iが決定論的動的治療レジメン$\bar{a}^g_K$を取ったときには 1 を，そうでないときには 0 を取る指示関数（indicator function）である．重みに関しては上記の $W^{\bar{A}}$ の他に，分子に治療に関する条件付き確率を取り入れた stabilized weight （安定化重み）W^{st}というものも存在する.[4, 5] 点治療に対して IPTW 法を行う際にはW^{st}を用いることは一般的であるが， 時間依存性治療に対しては注意が必要である．静的治療レジメンに対してはW^{st}を用いた IPTW 推定量は g-formula と同一の結果をもたらすが，動的治療レジメンに対しては g-formula とは異なる結果を導く．このため stabilized weight は動的治療レジメンでは一般的に用いられないことに注意されたい.

IPTW 推定量は共変量が離散変数であれば，過去の治療歴や時間依存性共変量（$\bar{A}_{t-1},\bar{L}_t$）が与えられた下でのt時点での治療に関する条件付き確率（分布）$f(A_t|\bar{A}_{t-1},\bar{L}_t)$ をデータに基づいて推定することによってノンパラメトリックに識別することができる．一方で時点数が長い場合やベースライン共変量，及び時間依存性共変量が高次元となる場合は，次元の呪いの影響を受ける．これを回避するためには治療確率に関してパラメトリックモデルを設定する必要があり，ロジスティック回帰モデルを用いることが通常である.

しかし，限られたサンプルサイズの下で，長い時点数にまたがった時間依存性治療の効果を推定するには，治療確率に関するパラメトリックモデルの仮定のみでは不十分である．前述のように，二値治療の治療時点がK個存在する場合，$E[Y^g]$は2^K個存在しサンプルサイズを大幅に超えるといったケースも珍しくない．この問題を解決する一つの方法は，$E[Y^g]$に対し制約を課すことである．周辺構造モデルは，まさにこの考えに基づくものであり，$E[Y^g]$を任意のモデル式で表現する.想定するモデル式は，研究における主要な目的に

よって異なり，加法モデルや乗法モデル，周辺構造平均モデルなど多種多様なモデル式が提案されている．ここでは具体例として，ヒト免疫不全ウイルス（HIV）感染患者に対して健康状態をスコア化したアウトカムをできるだけ高めることを目的とする，CD4 リンパ球数の数値を参照して抗レトロウイルス療法（cART）の実施する状況での周辺構造モデルのモデル式を考える.[5] 決定論的動的治療レジメンを $g = x$ で表し，その内容は「CD4 リンパ球数が x cell/μL 以下となった期間から継続して治療を続ける」と定義する．この時，ベースライン共変量の部分集合 V（共変量の値は二値）が与えられた下での周辺構造モデルが以下である．

$$E[\,Y^{g=x} \mid V\,] = h(x,V;\,\beta) = h_1(x,V;\beta_1) + h_2(V;\beta_2)$$

where $h_1(x,V;\beta_1) = \beta_{1,1}(x-350) + \beta_{1,2}(x-350)^2 + \beta_{1,3}(x-350)^3 + \beta_{1,4}(x-350)V, h_2(V;\beta_2) = \beta_{2,1} + \beta_{2,2}V$

ここで 350 という値は，「CD4 リンパ球数が 350 cell/μL 以下となった期間から継続して治療を続ける」という決定論的動的治療レジメンとの比較を考慮している．

治療確率に関するパラメトリックモデルと周辺構造モデルは，次元の呪いを回避するという目的は共通するものの，制約を置く対象がそれぞれ異なっているため，2 つのパラメトリックモデルは区別して考えられるべきである.[11] また，それぞれのモデルの誤特定はそれぞれ異なるバイアスを与えることにも注意が必要である．すなわち，治療確率に関するパラメトリックモデルを正しく特定出来ていたとしても周辺構造モデルを誤特定化している場合は IPTW 推定量の推定値にバイアスが発生する（逆の場合も同様)．この議論は IPTW 推定量を拡張して回帰モデル，あるいは治療確率モデルのいずれかが正しく特定できたのであれば，興味のある因果効果をバイアスなく推定可能となる二重頑健推定量（doubly robust estimators）に関しても同様に成立しており，それぞれのパラメトリックモデルがどの部分に対して制約を課しているかを認識することは非常に重要である．

決定論的動的治療レジメンにおける周辺構造モデルを用いた IPTW 法の他の留意点としては，ナイーブな周辺構造モデル（e.g.，治療の値の累積，cum($\bar{a}$)を利用したモデル）は動的治療レジメンに適合しないという点である.[12] これは動的治療レジメンが時間依存性治療のみならず，時間依存性共変量に基づいて治療が決定されるという特性上，その効果を推定するにあたっては時間依存性治療と時間依存性共変量の交互作用をモデリングする必要がある点に起因する．ナイーブな周辺構造モデルは，この交互作用のモデリングができないため，交互作用を考慮するにあたっては構造ネストモデルが用いられることが多い．構造ネストモデルについては本章冒頭の文献で概説されているため，適宜そちらを参照されたい．

最後に g-formula に対する IPTW 推定量の利用について述べる．本節の冒頭で触れたとおり，g-formula と IPTW 推定量は同じ推定対象 $E[Y^g]$ を識別している．これはどちらの方法を用いても理論的には同一の結果を与えることを意味しているため，治療時点数が多い，あるいは共変量が高次元ゆえにパラメトリックモデルを用いた場合でも parametric g-formula から推定された値とパラメトリックモデルを利用した IPTW 法の推定値を比較することは一定の意義がある．仮に 2 つの推定値間でサンプリングのばらつきで説明できる以上に差異が存在する場合，識別条件が成立するかにかかわらず，それぞれの方法で用いたパラメトリックモデルを誤特定していることが示唆される．そのため，2 つの推定値の差異を測ることでパラメトリックモデルを修正すべきか否かを検討することができ，結果として parametric g-formula の推定値の妥当性を高めることができる．ただし，2 つの推定値が同一であることはモデルを誤特定していないことを意味しない点には留意すべきである．

5. 結論 / SAS での実装

本稿では時間依存性治療と呼ばれる治療に対する統計的因果推論の概念，及びその因果効果の推定手法である g-methods の基本的な説明を行った．各手法は実装上の難易度，解釈のしやすさ，推定精度など様々な点で違いがあるが，それぞれの利点と欠点（注意点）を十分に理解した上で用いることが重要である．g-formula と 周辺構造モデルにおける IPTW 法は，$E[Y^g]$という同じ対象を推定しているため，実証上は parametric g-formula による推定値は パラメトリックモデルを利用した IPTW 法の推定値と比較されるべきである．仮に 2 つの推定値が同一であったとしてもモデルを誤特定化していないとは限らないが，推定値の比較はパラメトリックモデルを修正すべきかどうかの示唆を得られるため，結果の信頼性を高める観点からも比較の実施が推奨される．

周辺構造モデルにおける IPTW 法，並びに g-formula の SAS プログラムに関しては，著者の GitHub（https://github.com/Norihiro-Suzuki/SAS-Causal-inference）にて公開を行う．また，両手法のシミュレーションベースでの比較結果についても同様に上記の GitHub において一部公開する．

6. 参考文献

1. Pearl J. An introduction to causal inference. *Int J Biostat*. 2010;6(2):7.
2. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*. 1923;1-51.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.
4. Robins J, Hernan M. Estimation of the causal effects of time-varying exposures. Boca Raton; *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*; 2008. 553-599.
5. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton; *Chapman & Hall/CRC*, 2020.
6. 日本製薬工業協会. ICH E9(R1)の理解に役立つ因果推論～時間依存性治療編～. https://www.jpma.or.jp/information/evaluation/results/allotment/DS_202209_causal-tv.html. Accessed August 31, 2023.
7. Holland PW, Thayer DT. Differential item functioning and the Mantel-Haenszel procedure. *ETS Res Rep Ser*. 1986;2:1-24.
8. Young, JG, Hernán, MA, Robins, JM. Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiol Methods*. 2014;3:1-19.
9. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9-12):1393-1512.
10. McGrath S, Young JG, Hernán MA. Revisiting the g-null Paradox. *Epidemiology*. 2022;33:114-120.
11. Shinozaki T, Suzuki E. Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips. *J Epidemiol*. 2020 5;30(9):377-389.
12. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol*. 2006;98(3):237-242.

解析プロシジャで作成されるODS統計解析Plotについて

○折村奈美

（イーピーエス株式会社）

Creating Statistical Graphics with ODS in SAS

Nami Orimura

EPS Corporation

要旨

Output Delivery System (ODS)はプロシジャによって作成された表やグラフの出力を管理し, HTML や PDF など様々な形式で出力できる機能である. 本発表で取り上げる ODS 統計グラフは ODS の拡張機能である ODS Graphics により作成される. ODS Graphics を使用すると解析結果の作成と同時に自動的にグラフを作成することができ, 解析結果を解釈する際の補助として有用である. しかし, 統計解析業務の実務的場面においては解析結果のレポーティングそのものに優先的に労力が割かれるため, 作業時にグラフを有効活用して作業者自身で解析結果の考察や解釈を行うことができていない場面がある. そこで本発表では代表的なプロシジャについて例を交えながらこの機能と図の見方を説明し, その有用性を伝えることを目的とする.

キーワード: Output Delivery System, ODS Graphics, ODS 統計グラフ

1. 緒言

Output Delivery System と ODS Graphics

Output Delivery System (ODS) はプロシジャによって作成された表やグラフの出力を管理し, HTML や PDF など様々な形式で出力する機能である. ODS を使用することで以下のことが可能になり, 出力の作成, 保存をより柔軟に行うことができる.

・一般的なアプリケーション用のレポートの作成

SAS 以外のソフトウェア専用の出力を作成できる. 例えば ODS PDF ステートメントを使用して Adobe Acrobat で表示することや印刷用の PDF ファイルを作成することができる. PDF のほかに HTML, RTF, EPUB などの形式でも出力可能である.

・レポート内容のカスタマイズ

グラフィックを埋め込むこと, 特定のセル内容を選択して表示すること, 表やグラフに埋め込みリンクを作成することが可能である. また, プロシジャ出力から特定の表やグラフを選択または除外して印刷することも可能である.

・体裁のカスタマイズ

出力の色, フォント, レイアウト, ヘッダーの変更, 画像や埋め込み URL の追加などが可能である.

ODS Graphics は ODS の拡張機能の一つで，解析プロシジャにおいてグラフを作成する機能である．SAS9.4 で ODS Graphics を使用できるプロシジャを以下の Table 1 に示す．この機能は解析結果を解釈する際の補助として有用であるが，統計解析業務の実務的場面においては解析結果のレポーティングそのものに優先的に労力が割かれるため，作業時にグラフを有効活用して作業者自身で解析結果の考察や解釈を行うことができていない場面がある．そこで本稿では代表的なプロシジャについて ODS Graphics の機能を説明し，その有用性を示すことを目的とする．

Table 1. SAS 9.4 における ODS Graphics をサポートする統計プロシジャ

STATISTICAL PROCEDURES THAT SUPPORT ODS GRAPHICS IN SAS 9.4

The following statistical procedures support ODS Graphics in SAS 9.4:

SAS/STAT		**Base SAS**	**SAS/ETS**
ACECLUS	MIXED	CORR	ARIMA
ADAPTIVEREG	MULTTEST	FREQ	AUTOREG
ANOVA	NLIN	UNIVARIATE	COPULA
BCHOICE	NPAR1WAY		COUNTREG
BOXPLOT	ORTHOREG	**SAS/QC**	ENTROPY
CALIS	PHREG	ANOM	ESM
CLUSTER	PLM	CAPABILITY	EXPAND
CORRESP	PLS	CUSUM	HPCDM
FACTOR	POWER	MACONTROL	HPQLIM
FMM	PRINCOMP	MVPDIAGNOSE	HPSEVERITY
FREQ	PRINQUAL	MVPMONITOR	MODEL
GAM	PROBIT	MVPMODEL	PANEL
GAMPL	QUANTLIFE	PARETO	PDLREG
GEE	QUANTREG	RAREEVENTS	QLIM
GENMOD	QUANTSELECT	RELIABILITY	SEVERITY
GLIMMIX	REG	SHEWHART	SIMILARITY
GLM	ROBUSTREG		SSM
GLMPOWER	RSREG	**Other**	SYSLIN
GLMSELECT	SEQDESIGN	HPF	TIMEDATA
HPFMM	SEQTEST	HPFENGINE	TIMEID
HPSPLIT	SIM2D		TIMESERIES
ICLIFETEST	SPP	SAS Risk	UCM
ICPHREG	STDRATE	Dimensions	VARMAX
IRT	SURVEYFREQ		X12
KDE	SURVEYLOGISTIC		
KRIGE2D	SURVEYMEANS		
LIFEREG	SURVEYPHREG		
LIFETEST	SURVEYREG		
LOESS	TPSPLINE		
LOGISTIC	TRANSREG		
MCMC	TTEST		
MDS	VARCLUS		
MI	VARIOGRAM		

ODS Graphics を用いてグラフを作成する際には事前に ODS Graphics を有効にする必要がある．メニューバーのツール→オプション→プリファレンスをクリックすると Fig 1 のウィンドウが表示されるので，“ODS Graphics を使用する”にチェックを入れる．または以下のコードをプロシジャの前に記載することで ODS Graphics の機能を有効化できる．

```
ods graphics on;
```

さらに“HTML を作成する”にチェックを入れることで結果が自動的に HTML 形式で表示され，グラフを確認する際に有用である．こちらも ODS Graphics の有効化と同様に以下のコードをプロシジャの前に記載することでも有効化できる．

```
ods html;
```

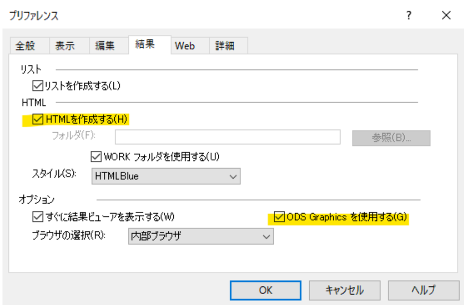


Fig 1. SAS のプリファレンスウィンドウ

2. グラフ名の取得と SAS リファレンスを用いたグラフ解釈の調べ方

REG プロシジャで出力されるグラフセットを例に，各グラフ名の取得方法を説明する．まず結果ツリーを右クリックしてプロパティを開き，グラフセット名を取得する(Fig 2. b, c)．その後プロシジャのオプションで“plots(only)=グラフセット名(unpack)”と指定することでグラフセットを解体し，結果ツリーから各グラフの名前を取得する(Fig 2. d).

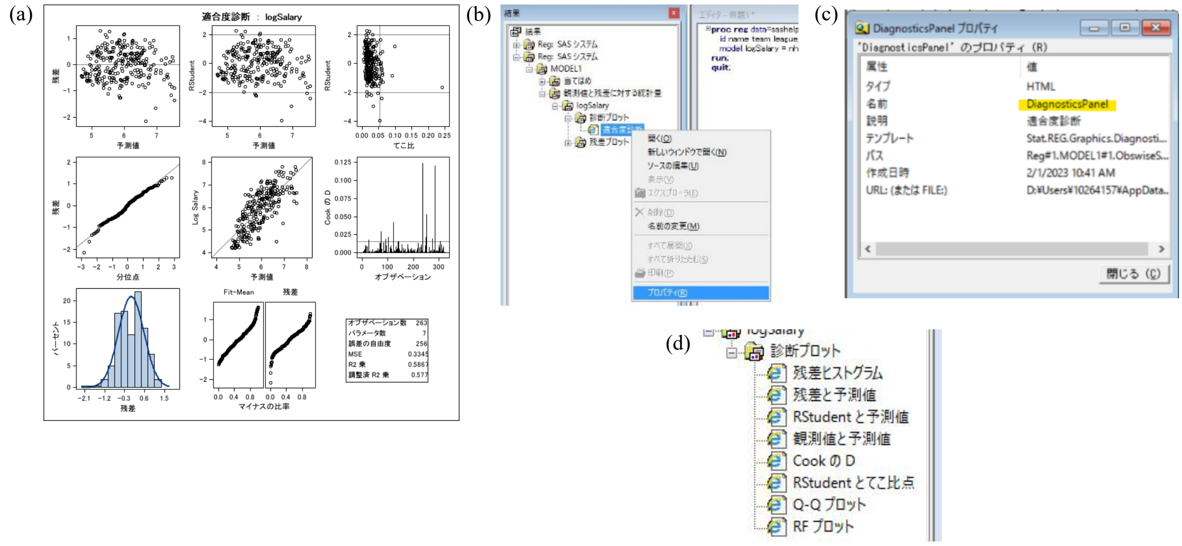


Fig 2. グラフ名の取得手順

```
proc reg data=sashelp.baseball plots(only)=DiagnosticsPanel(unpack);
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
quit;
```

グラフの名前を取得できればグラフの詳細について調べることができる. SAS リファレンスでプロシジャを検索し, ODS Graphics の項目を見ると解説や引用元の論文について記載がある(Fig 3. a). さらにサンプルグラフのリンクが記載されているプロシジャもあり，そのリンクをクリックすると SAS 社によるグラフの解釈を確認できる(Fig 3. b).

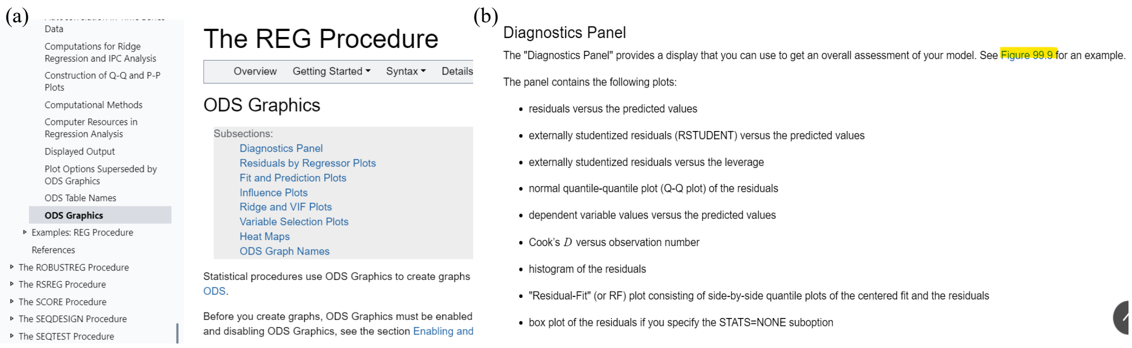


Fig 3. SAS リファレンスの ODS Graphics の解説

3. FREQ プロシジャ

3.1 FreqPlot と CumfreqPlot

FREQ プロシジャで作成される度数プロット，累積度数プロットについて説明する. sashelp ライブラリに格納されている class データセットを使用して，年齢のプロットを作成することとする. tables ステートメントのオプションで“plots=(freqplot cumfreqplot)”と指定して実行すると以下の通り Table 2 の度数列が Fig 4. a，累積度数列が Fig 4. b のグラフとして出力される．このようにオプションを指定するだけで度数プロットや累積度数プロットを作成でき，データの特徴をとらえるうえで有用である.

```
proc freq data=sashelp.class;
   tables age/ plots=(freqplot cumfreqplot);
run;
```

Table 2. 実行結果

年齢				
Age	度数	パーセント	累積度数	累積パーセント
11	2	10.53	2	10.53
12	5	26.32	7	36.84
13	3	15.79	10	52.63
14	4	21.05	14	73.68
15	4	21.05	18	94.74
16	1	5.26	19	100.00

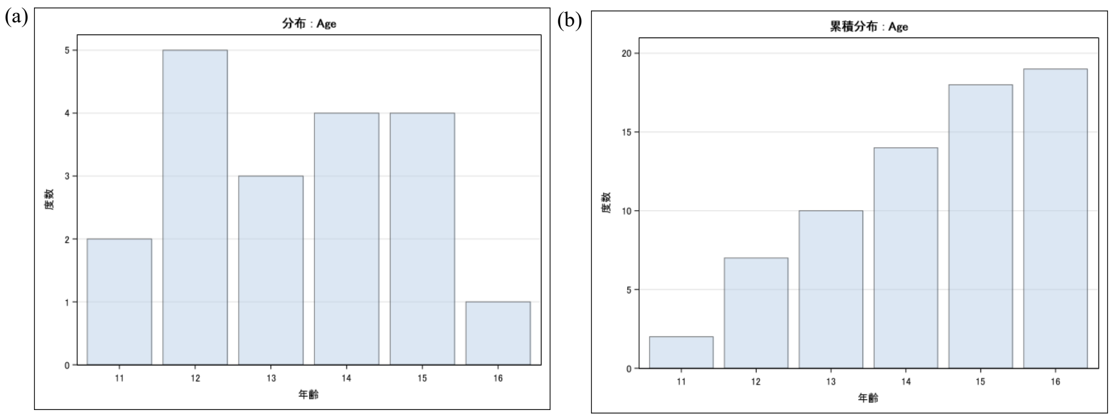


Fig 4. FreqPlot (a)および CumfreqPlot (b)

3.2 MosaicPlot

以下のテストデータを作成し, tables ステートメントのオプションで“plots=mosaicplot”と指定することで Fig 5 のプロットを描画できる．グラフの横幅はカテゴリーの各層の度数の合計に比例しており，この場合 TRTAN=1 が 6 例, TRTAN=2 が 3 例であるので横幅の比は 2:1 となっている．また、四角形の高さはカテゴリーの各層の度数に比例しており, TRTAN=1 では上から 2:1:1:2, TRTAN=2 では上から 1:1:1 となっている．このように群ごとの数の偏りや群内での比率を可視化できるので，合併症・既往歴の SOC 発現や有害事象などの確認に適している.

```
data test;
length AEDECOD $200;
   TRTAN=1;
   AEDECOD="Rhinitis"; output;
   AEDECOD="Diarrhea"; output;
   AEDECOD="Nausea"; output;
   AEDECOD="Sleepiness"; output;
   AEDECOD="Diarrhea"; output;
   AEDECOD="Sleepiness"; output;
   TRTAN=2;
   AEDECOD="Sleepiness"; output;
   AEDECOD="Nausea"; output;
   AEDECOD="Fever"; output;
run;

proc freq data=test;
   tables AEDECOD*TRTAN/ plots=mosaicplot;
run;
```

Table 3. 実行結果

度数 パーセント 行のパーセント 列のパーセント	表：AEDECOD * TRTAN		
	TRTAN		
AEDECOD	1	2	合計
Diarrhea	2 22.22 100.00 33.33	0 0.00 0.00 0.00	2 22.22
Fever	0 0.00 0.00 0.00	1 11.11 100.00 33.33	1 11.11
Nausea	1 11.11 50.00 16.67	1 11.11 50.00 33.33	2 22.22
Rhinitis	1 11.11 100.00 16.67	0 0.00 0.00 0.00	1 11.11
Sleepiness	2 22.22 66.67 33.33	1 11.11 33.33 33.33	3 33.33
合計	6 66.67	3 33.33	9 100.00

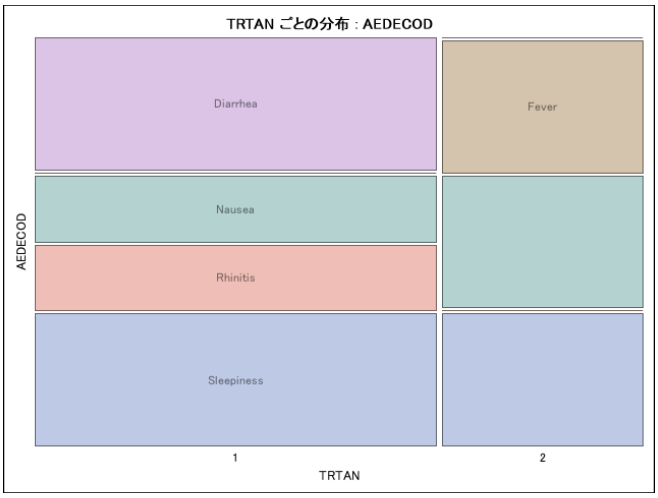


Fig 5. MosaicPlot

3.3 リスク・オッズ

クロス集計表のデータが GROUP=1, 2 ごとにあるとして, tables ステートメントのオプションで“OR”, “RELRISK”, “RISKDIFF”と指定することでそれぞれオッズ比, 相対リスク指標と信頼限界, リスク・合計リスク・リスク差が計算される. “plots(COLUMN=1)”で列 1（N 列）のグラフを描くことを指定して, “ODDSRATIOPLOT”, “RELRISKPLOT”, “RISKDIFFPLOT”でそれぞれの統計量に対応するグラフ名を指定する. “CL=SCORE”で Wilson's score 法の信頼限界を, “CL=WALD”で Wald 法の信頼限界を算出し, “STATS”でグラフに統計量を表示する. このコードを実行すると以下の結果が出力される(Table 4, Fig 6). Table 4 に GROUP=1 の統計量のみ掲載したが, GROUP=2 の統計量についても同様に出力される. このグラフからリスク比やオッズ比の群間差を可視化できるので, サブグループ解析の QC などに有用である.

```
data test;
  GROUP=1;
    A="Y";B="Y";_FREQ_=12;output;
    A="Y";B="N";_FREQ_=8;output;
    A="N";B="Y";_FREQ_=15;output;
    A="N";B="N";_FREQ_=5;output;
  GROUP=2;
    A="Y";B="Y";_FREQ_=11;output;
    A="Y";B="N";_FREQ_=10;output;
    A="N";B="Y";_FREQ_=8;output;
    A="N";B="N";_FREQ_=12;output;
run;

proc freq data=test;
  tables GROUP * A * B / OR RELRISK RISKDIFF
    plots(COLUMN=1)= ODDSRATIOPLOT (CL=SCORE STATS)
    plots(COLUMN=1)= RELRISKPLOT (CL=WALD STATS)
    plots(COLUMN=1)= RISKDIFFPLOT (STATS)
;
  weight _FREQ_ /zeros;
run;
```

Table 4. 実行結果（GROUP=1 について）

度数 パーセント 行のパーセント 列のパーセント

表 1 : A * B
層別変数 : GROUP=1

A	B: N	B: Y	合計
N	5 12.50 25.00 38.46	15 37.50 75.00 55.56	20 50.00
Y	8 20.00 40.00 61.54	12 30.00 60.00 44.44	20 50.00
合計	13 32.50	27 67.50	40 100.00

表 1 : A * B の統計量
層別変数 : GROUP=1

列 1 リスクの推定値

	リスク	ASE	95% 信頼限界		正確 95% 信頼限界	
行 1	0.2500	0.0968	0.0602	0.4398	0.0866	0.4910
行 2	0.4000	0.1095	0.1853	0.6147	0.1912	0.6395
合計	0.3250	0.0741	0.1799	0.4701	0.1857	0.4913
差	−0.1500	0.1462	−0.4366	0.1366		

行 1 − 行 2 の差

列 2 リスクの推定値

	リスク	ASE	95% 信頼限界		正確 95% 信頼限界	
行 1	0.7500	0.0968	0.5602	0.9398	0.5090	0.9134
行 2	0.6000	0.1095	0.3853	0.8147	0.3605	0.8088
合計	0.6750	0.0741	0.5299	0.8201	0.5087	0.8143
差	0.1500	0.1462	−0.1366	0.4366		

行 1 − 行 2 の差

オッズ比と相対リスク

統計量	値	95% 信頼限界	
オッズ比	0.5000	0.1295	1.9303
相対リスク (列 1)	0.6250	0.2467	1.5836
相対リスク (列 2)	1.2500	0.8064	1.9375

オッズ比の信頼限界
オッズ比 = 0.5000

タイプ	95% 信頼限界	
スコア	0.1327	1.8974

標本サイズ = 40

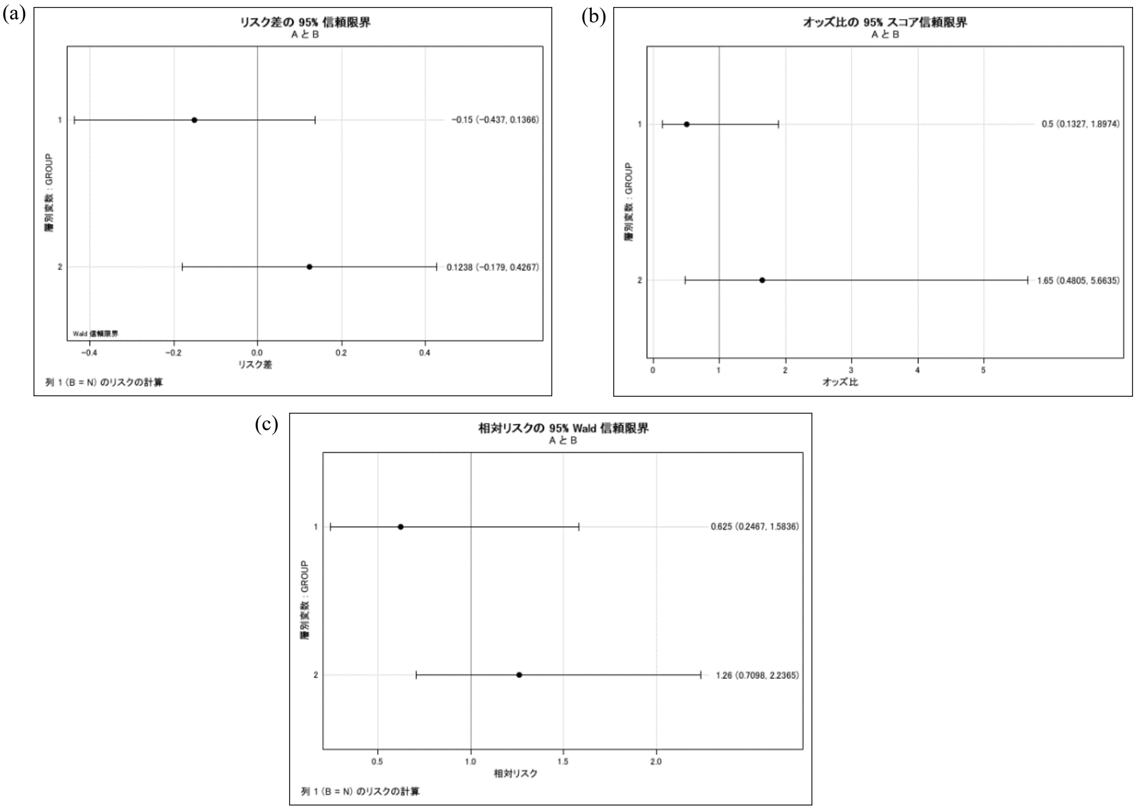


Fig 6. リスク・オッズのグラフ

3.4 AgreePlot と KappaPlot

tables ステートメントのオプションで“agree”と指定すると, McNemar 検定の結果と Kappa 係数, AgreePlot が出力される(Table 5, Fig 7). McNemar 検定は対応のあるペアの 2 値データを用いて 2 つの処理の結果に差があるかどうかを検定する手法で, Kappa 係数は 1 に近づくほど 2 つの結果の一致度合いが高いことを示す指標である. Table 5 から McNemar 検定の p 値が 0.0016 であるので Before と After の結果に有意差があり, Kappa 係数が 0.1500 であることから Before と After の一致度が低いことが読み取れる. さらに AgreePlot の濃青色の部分は Before と After の結果が同じだった人を示しており, 薄青色と濃青色の面積の差が大きいほど Before と After の一致度が低いことを表している. また, 45 度線と薄青色の四角形の交点のずれが全体に対しての比の偏り具合を示しており, このグラフでは四角形の交点よりも 45 度線が上部にあり, Y の割合が高いことを示している. もし Y と N の割合が等しければ薄青色の四角形は正方形となり, 角は 45 度線上に重なる. このように一致度と全体に対する比の偏り具合を可視化でき有用である.

```
data test;
   BEFORE="Y";AFTER="Y";_FREQ_=18;output;
   BEFORE="Y";AFTER="N";_FREQ_=2;output;
   BEFORE="N";AFTER="Y";_FREQ_=15;output;
   BEFORE="N";AFTER="N";_FREQ_=5;output;
run;

proc freq data=test;
   tables BEFORE * AFTER/ agree;
   weight _FREQ_/ zeros;
run;
```

Table 5. 実行結果

度数 パーセント 行のパーセント 列のパーセント	表 : BEFORE * AFTER			
		AFTER		
	BEFORE	N	Y	合計
	N	5 12.50 25.00 71.43	15 37.50 75.00 45.45	20 50.00
	Y	2 5.00 10.00 28.57	18 45.00 90.00 54.55	20 50.00
	合計	7 17.50	33 82.50	40 100.00

BEFORE * AFTER の統計量

McNemar の検定		
カイ 2 乗値	自由度	Pr > ChiSq
9.9412	1	0.0016

単純カッパ係数			
推定値	標準誤差	95% 信頼限界	
0.1500	0.1188	-0.0828	0.3828

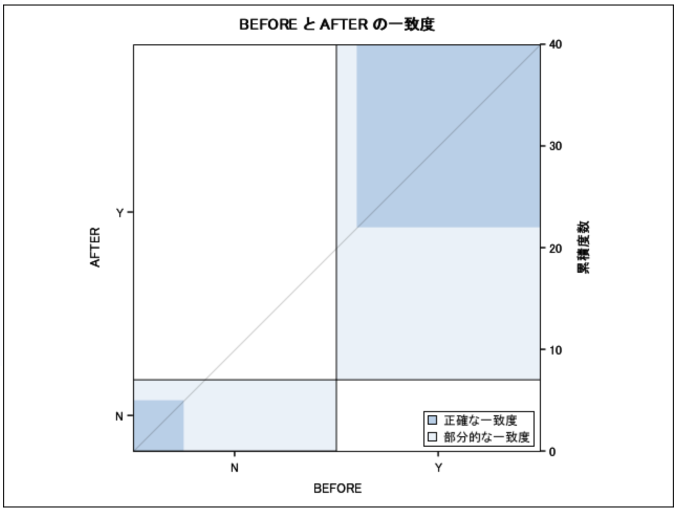


Fig 7. AgreePlot

また, GROUP=A, B ごとに Before と After の結果があるような多元平方表の場合, オプションを"agree plots=kappaplot"とすることで AgreePlot に加えて KappaPlot が作成される(Fig 8). GROUP=B および全体についても Table 6 のような表と統計量が出力され, カッパ係数と信頼区間が Fig 8 のグラフとして表示される. このグラフは群ごとに一致度を比較する際に活用できる.

```
data test;
  GROUP="A";
    BEFORE="Y";AFTER="Y";_FREQ_=12;output;
    BEFORE="Y";AFTER="N";_FREQ_=8;output;
    BEFORE="N";AFTER="Y";_FREQ_=15;output;
    BEFORE="N";AFTER="N";_FREQ_=5;output;

  GROUP="B";
    BEFORE="Y";AFTER="Y";_FREQ_=18;output;
    BEFORE="Y";AFTER="N";_FREQ_=2;output;
    BEFORE="N";AFTER="Y";_FREQ_=3;output;
    BEFORE="N";AFTER="N";_FREQ_=17;output;
run;

proc freq data=test;
    tables GROUP * BEFORE * AFTER/ agree plots=kappaplot;
    weight _FREQ_/ zeros;
run;
```

Table 6. 実行結果

度数 パーセント 行のパーセント 列のパーセント	表 1 : BEFORE * AFTER		
	層別変数 : GROUP=A		
	AFTER		
BEFORE	N	Y	合計
N	5 12.50 25.00 38.46	15 37.50 75.00 55.56	20 50.00
Y	8 20.00 40.00 61.54	12 30.00 60.00 44.44	20 50.00
合計	13 32.50	27 67.50	40 100.00

表 1 : BEFORE * AFTER の統計量
層別変数 : GROUP=A

McNemar の検定		
カイ 2 乗値	自由度	Pr > ChiSq
2.1304	1	0.1444

単純カッパ係数			
推定値	標準誤差	95% 信頼限界	
-0.1500	0.1464	-0.4370	0.1370

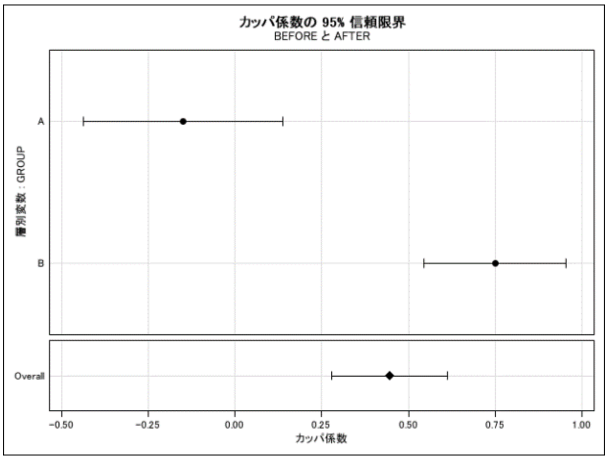


Fig 8. KappaPlot

4. CORR プロシジャ

4.1 回帰分析

2.1 と同様の class データセットを使用して身長と体重の相関関係を評価することとする. corr プロシジャのオプションで"plots=scatter"と指定することで相関係数を算出できる. 以下のコードを実行した結果, 単純統計量とピアソンの相関係数および散布図が出力される(Table 7, Fig 9). 身長と体重の相関係数が 0.87779 と強い正の相関があることが示され, 散布図が右肩上がりであることからも強い正の相関があることが読み取れる. 散布図中の青色の楕円は予測楕円である. これは母集団から抽出された新たなデータがとり得る範囲をさし, Fig 9 の 95%予測楕円は母集団から新しく 100 回データを抽出した際に 95 回がこの楕円内にプロットされるであろうことを示している. このように出力される散布図を活用することで相関関係を視覚的にとらえることができる.

```
proc corr data=sashelp.class plots=scatter;
var Height Weight;
run;
```

Table 7. 実行結果

CORR プロシジャ

2 変数 :	Height Weight

単純統計量							
変数	N	平均	標準偏差	合計	最小値	最大値	ラベル
Height	19	62.33684	5.12708	1184	51.30000	72.00000	身長(インチ)
Weight	19	100.02632	22.77393	1901	50.50000	150.00000	体重(ポンド)

Pearson の相関係数, N = 19
H0: Rho=0 に対する Prob > |r|

	Height	Weight
Height 身長(インチ)	1.00000	0.87779 <.0001
Weight 体重(ポンド)	0.87779 <.0001	1.00000

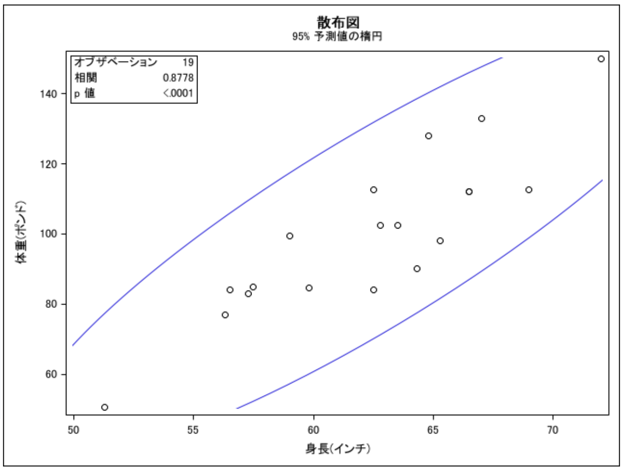


Fig 9. 散布図

さらに解析対象変数が 3 つ以上あるときは散布図行列を作成できる．先ほどのコードの解析対象変数に Age を追加して，“plots=scatter”を“plots=matrix(histogram)”に変更する．このコードを実行した結果，全組み合わせの相関係数と散布図行列が出力される(Table 8, Fig 10)．このグラフを活用することで相関がある変数の組み合わせを視覚的にとらえることができ有用である．

```
proc corr data=sashelp.class plots=matrix(histogram);
   var Height Weight Age;
run;
```

Table 8. 実行結果

CORR プロシジャ

3 変数：	Height Weight Age

単純統計量

変数	N	平均	標準偏差	合計	最小値	最大値	ラベル
Height	19	62.33684	5.12708	1184	51.30000	72.00000	身長(インチ)
Weight	19	100.02632	22.77393	1901	50.50000	150.00000	体重(ポンド)
Age	19	13.31579	1.49267	253.00000	11.00000	16.00000	年齢

Pearson の相関係数, N = 19
H0: Rho=0 に対する Prob > |r|

	Height	Weight	Age
Height 身長(インチ)	1.00000	0.87779 <.0001	0.81143 <.0001
Weight 体重(ポンド)	0.87779 <.0001	1.00000	0.74089 0.0003
Age 年齢	0.81143 <.0001	0.74089 0.0003	1.00000

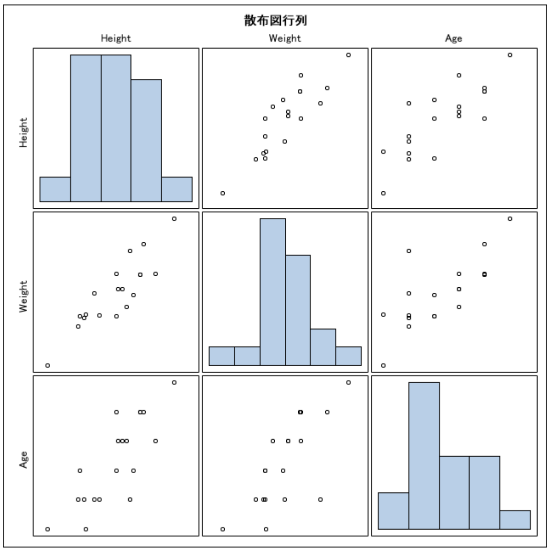


Fig 10. 散布図行列

5. REG プロシジャ

5.1 回帰分析

sashelp ライブラリの baseball データセットを使用して，年俸の対数を目的変数として，ヒット数，出塁数，打点，四球出塁数，メジャー年数，通算安打を説明変数として回帰分析を行う．以下のコードを実行した結果，Table 9, Fig 11, Fig 13 が出力される．Table 9 のパラメータ推定値から，以下の回帰式を求めることができる．

$$Y(logSalary)=4.1461+0.0066X_{nHits}+0.0002X_{nRuns}+0.0013X_{nRBI}+0.0067X_{nBB}+0.0711X_{Yrmajor}+0.0002X_{CrHits}$$

この回帰モデルの適合度をはかる指標として，決定係数が使用される．重回帰分析の場合決定係数として調整済み R2 乗値を採用することとなっており，この例では 0.5770 である．しかしモデルを決定係数のみで評価するのは必ずしも適切ではなく，残差の分布も確認したうえで評価することが望ましいとされている．その際に ODS Graphics で出力されるグラフを活用する．

```
proc reg data=sashelp.baseball;
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
quit;
```

Table 9. 実行結果

REG プロシジャ
モデル：MODEL1
従属変数：logSalary Log Salary

読み込んだオブザベーション数	322
使用されたオブザベーション数	263
欠損値を含むオブザベーション数	59

分散分析

要因	自由度	平方和	平均平方	F 値	Pr > F
Model	6	121.53052	20.25509	60.56	<.0001
Error	256	85.62322	0.33447		
Corrected Total	262	207.15373			

Root MSE	0.57833	R2 乗	0.5867
従属変数の平均	5.92722	調整済み R2 乗	0.5770
変動係数	9.75719		

パラメータの推定

変数	ラベル	自由度	パラメータ推定値	標準誤差	t 値	Pr > |t|
Intercept	Intercept	1	4.14614	0.13612	30.46	<.0001
nHits	Hits in 1986	1	0.00663	0.00210	3.15	0.0018
nRuns	Runs in 1986	1	0.00019890	0.00398	0.05	0.9602
nRBI	RBIs in 1986	1	0.00125	0.00235	0.53	0.5947
nBB	Walks in 1986	1	0.00672	0.00239	2.81	0.0054
YrMajor	Years in the Major Leagues	1	0.07108	0.01925	3.69	0.0003
CrHits	Career Hits	1	0.00023910	0.00014571	1.64	0.1020

REG プロシジャで回帰分析を行った際に出力されるグラフセット（Fig 11）について，各グラフの見方を説明する．

(a)のグラフは残差と予測値のグラフである．残差が縦軸 0 に対して均一に分布している場合はモデルが適しており，プロットに何か傾向がみられる場合，そのモデルは妥当ではないと判断される．この場合グラフ右下に外れ値がいくつかみられるものの残差はバランスよく散らばっており，モデルは概ね妥当であると考えられる．

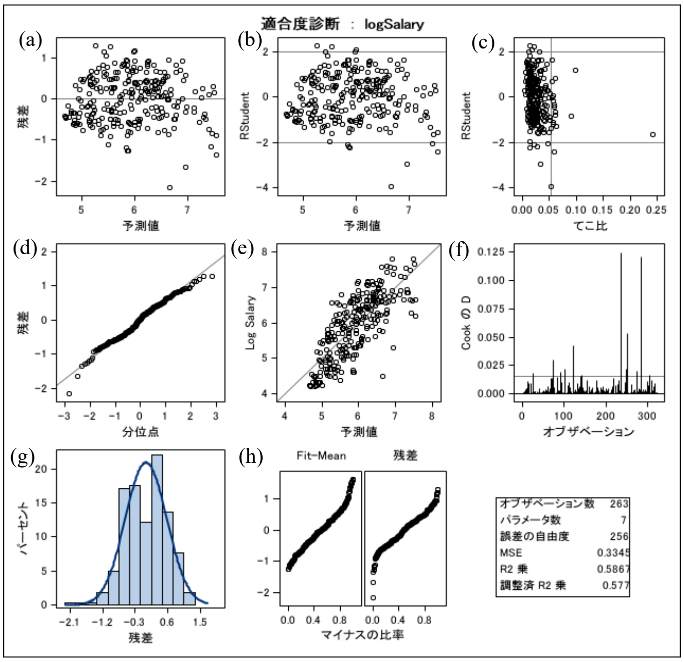


Fig 11. REG プロシジャで出力されるグラフセット

(b)のグラフは Rstudent と予測値のグラフである．Rstudent はスチューデント化残差を指し，残差を補正標準誤差で割って補正をかけた値で，絶対値が 2 より大きいデータを外れ値とみなす．このグラフでは Rstudent が絶対値 2 の間にバランスよく散らばっているので，モデルは概ね妥当であると考えられる．

(c)のグラフは Rstudent とてこ比のグラフで，モデルの適合度に影響を与えるデータを見つけやすいグラフである．てこ比はデータが全体の平均からどの程度ずれているかを示す値で，縦線はてこ比が 2(p+1)/n（※1）のラインを示しており，このラインを超えると影響力の高いデータであるとみなされる．このグラフからてこ比が極端に大きいデータが一つあり，これがモデルの適合度に大きな影響を与えていると考えられる．

※1: p=パラメータ数, n=観測データ数

(d)は正規分布かどうかを確認する Q-Q プロットで，この場合殆どのプロットが直線に重なっているので残差はほぼ正規分布しているといえる．

(e)は観測値と予測値のグラフで, 45 度線からのばらつきでモデルの適合度が可視化されたグラフである．このグラフからモデルが目的変数の動作をおおよそ予測していて適合度が高いと考えられる．

(f)のグラフは Cook の距離のグラフである．0 と 0.025 の間に引かれた直線は Cook の距離が 4/n のラインであり，このラインを超えているときにモデルに対して影響力のあるデータであるとみなされる．このグラフ

から非常に影響力の高いデータが 2 つあることが読み取れる．

(g)は残差ヒストグラムで，残差がほぼ正規分布していることが示されている．

(h)は RF プロットである．左の Fit-Mean のグラフは予測値から目的変数の平均値を引いた値を，右は残差をプロットして分布を比較しているグラフである．これは左右のグラフの縦方向の広がり具合を比較するもので，左のプロットよりも右のプロットの広がりが小さい場合，モデルは妥当であると判断される．この結果から右図の左下が外れ値であり，それを除くと左図よりも右図の縦の広がりが小さいため，モデルは妥当であると考えられる．

(c)と(f)のグラフからモデルに影響を与えるデータが明らかになったので，このようなデータを特定するためにグラフにラベルを付けることとする．reg プロシジャのオプションで “plots(only label)=(RStudentByLeverage CooksD)”として(c)と(f)のグラフ名を指定すると，Fig 12 が出力される．このようにグラフにラベルを付けることで特異な値を速やかに特定することができ有用である．

```
proc reg data=sashelp.baseball plots(only label)=(RStudentByLeverage CooksD);
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
quit;
```

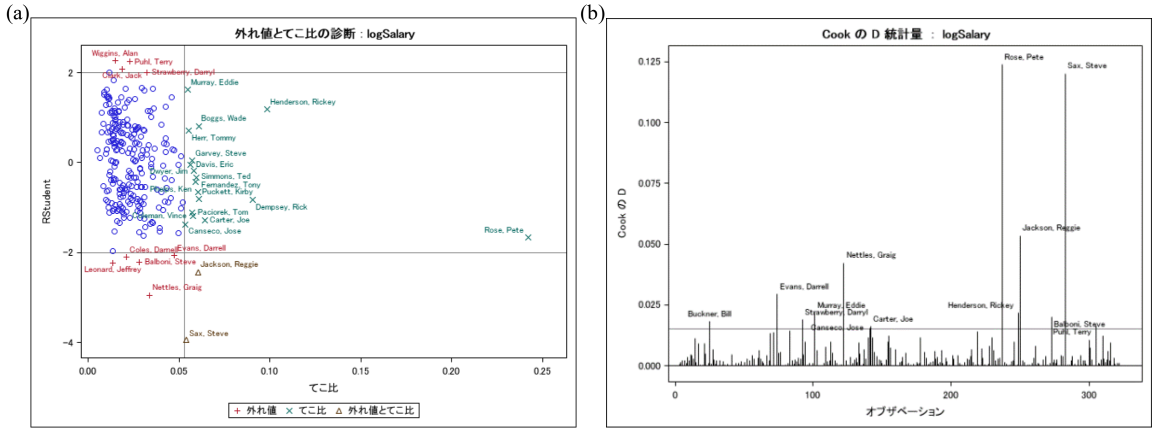


Fig 12. ラベル付けしたグラフ

さらに，Fig 11 とともに出力される Fig 13 は残差と回帰子（説明変数）のグラフで，ヒット数，出塁数などパラメータごとに分けて作成されている．デフォルトではグラフのパターンを表す曲線は表示されないが，reg プロシジャのオプションで “plots=residuals(smooth)”と指定することで表示できる．このグラフより，メジャー在籍年数が短い人と長い人はモデルの予測よりも報酬が少なく，モデルによる予測の精度が高くないことが読み取れる．このように条件によって分けて残差を確認することで，モデルを多角的に評価することができる．

以上のように，決定係数だけでなく残差を確認することで得られたモデルが適切ではないときに気付くことができ，その際に ODS Graphics で描かれるグラフが活用される．

```
proc reg data=sashelp.baseball plots=residuals(smooth);
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
quit;
```

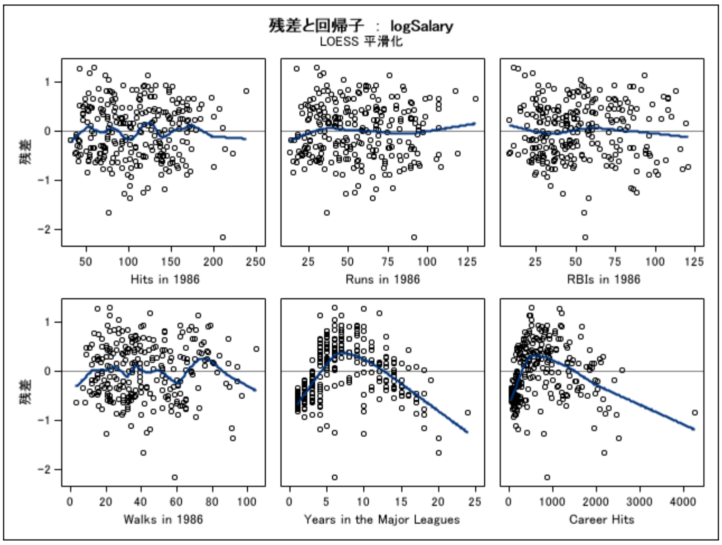


Fig 13. 残差と説明変数のグラフ

6. LIFETEST プロシジャ

6.1 Kaplan-Meier 曲線

骨髄移植患者の無病生存期間を格納した BMT データセットを使用して Kaplan-Meier 曲線を作成することとする．以下のコードを実行した結果, Fig 14 のグラフが出力される．“plots=survival”で生存曲線を描くことを指定し, “atrisk~”の部分で Number at Risk を軸の外側に出力することを指定する．さらに “cb=hw”で Hall-Wellner の信頼区間を表示し, “test”と“test=wilcoxon”で Wilcoxon 検定の結果を表示できる．臨床試験の統計解析業務ではデータセットを ods output で取り出して SGPLOT で描画することが多いが, 本項で示すように LIFETEST プロシジャでも Kaplan-Meier 曲線を作成でき，オプションを工夫することで出力をコントロールすることができる．

```
proc lifetest data=sashelp.BMT
        plots=survival (atrisk(outside)=0 to 3000 by 500 cb=hw test);
    where group ne "ALL";
    time t * status(0);
    strata group/ test=wilcoxon;
run;
```

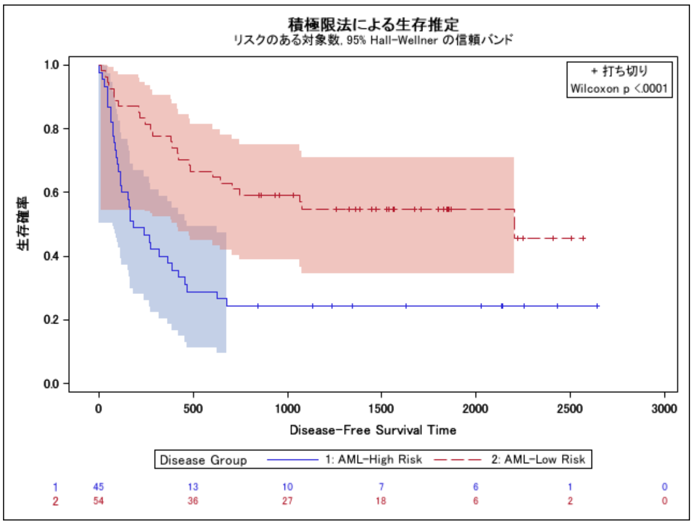


Fig 14. Kaplan-Meier 曲線

6.2 log-log 生存率グラフと RMST Plot

LIFETEST プロシジャの plots オプションで“loglogs”と指定すると log-log 生存率グラフが出力される(Fig 15. a). AML-High Risk と AML-Low Risk のように曲線が交差せず平行に近ければ比例ハザード性が成立していると判断できる．このグラフは log-rank 検定や比例ハザードモデルに基づく尤度比検定など，比例ハザード性の成立が前提となる解析を行う際に活用できる．

さらに plots オプションで“rmst”と指定することで，境界時間 tau を 0 から最終観測時点までスライドさせた際の境界内平均生存時間（RMST）の変動を示すグラフが出力される(Fig 15. b). RMST は比例ハザード性に依存しない評価指標として用いられることが多い．このグラフは同一の薬剤を用いた別試験の計画などの際に，境界時間をどのように設定するか検討する場面で有用である．

```
proc lifetest data=sashelp.BMT plots=(loglogs rmst);
   time t * status(0);
   strata /group=GROUP;
run;
```

(a) (b)

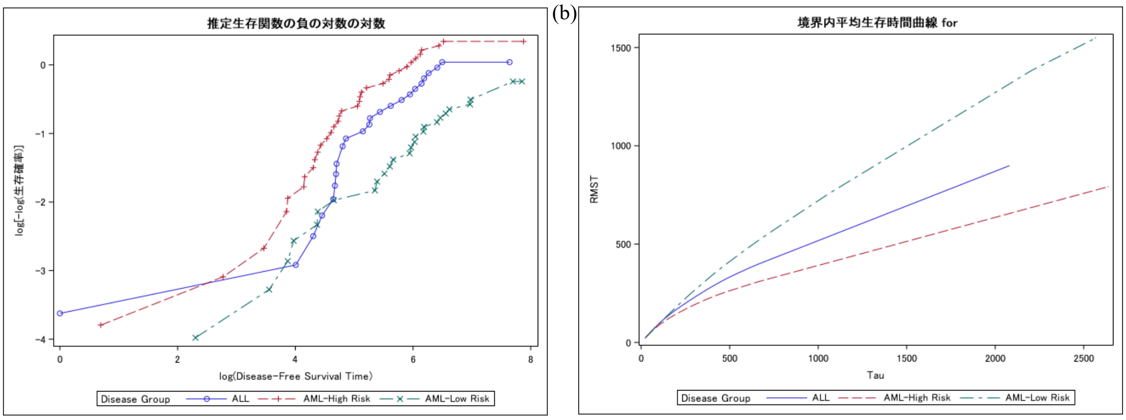


Fig 15. log-log 生存率グラフ(a)と RMST Plot(b)

7. まとめ

FREQ, CORR, REG, LIFETEST プロシジャを例に ODS Graphics の機能とグラフの解釈について紹介した．ODS Graphics で出力されるグラフは解析結果の評価・判断やその効率化のために活用できること，表からは読み取れなかったパターンや差を読み取れることから非常に有用であると確認できた．グラフの出力に複雑なコードを必要とせず，簡単なオプションを指定するだけでよい点も大きなメリットである．また，SAS リファレンスからグラフの詳細や解釈例を確認することで，数式とは異なる視点から解析手法について理解を深めることも可能である．今後 ODS Graphics の有用性が広く認識され，活用される機会が増えれば幸いである．

8. 参考文献

[1] SAS Help Center, “SAS/STAT User’s Guide Using the Output Delivery System”, 2021, p578-650

[2] SAS Help Center, “SAS 9.4 and SAS Viya 3.5 Programming Documentation | SAS 9.4/Viya 3.5 | Output and Graphics”, https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/pgmsasrptwlcm/home.htm, 2021, Accessed

Aug 15, 2023

[3] Robert N. Rodriguez and Warren F. Kuhfeld, SAS Institute Inc., Cary, NC, “An Overview of ODS Statistical Graphics in SAS® 9.4”, 2011

[4] SAS Help Center, “SAS 9.4 and SAS Viya 3.5 Programming Documentation | SAS 9.4/Viya 3.5 | The FREQ Procedure | ODS Graphics”, https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_freq_details125.htm, 2020, Accessed Aug 16, 2023

[5] Social Survey Research Information Co., Ltd., “2-2. モザイク図を描いてみよう｜統計学の時間｜統計WEB”, https://bellcurve.jp/statistics/course/18862.html, Accessed Aug 16, 2023

[6] Analyse-it Software, Ltd., “Agreement plot > Method comparison / Agreement > Statistical Reference Guide | Analyse-it® 6.10 documentation”, https://analyse-it.com/docs/user-guide/method-comparison/agreementplot, 2023, Accessed Aug 16, 2023

[7] Shrikant Bangdiwala, “The agreement chart as an alternative to the receiver-operating characteristic curve for diagnostic tests”, 2008, Journal of Clinical Epidemiology 61, p866-874

[8] SAS Institute, “SAS Help Center: ODS Graphics(REG)”, https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/statug/statug_reg_details53.htm, 2022, Accessed Aug 16, 2023

[9] Social Survey Research Information Co., Ltd., “予測値と残差|統計学の時間|統計 WEB”, https://bellcurve.jp/statistics/course/9704.html, Accessed Aug 16, 2023

[10] 株式会社アイスタット, “重回帰分析とは?重回帰分析の概要と具体例・結果”, https://istat.co.jp/ta_commentary/multiple_02, 2014, Accessed Aug 16, 2023

[11] SAS Institute, “SAS Help Center: ODS Graphics(LIFETEST)”, https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/statug/statug_lifetest_details86.htm, 2022, Accessed Aug 16, 2023

[12] SAS Institute, “Base SAS® 9.3 プロシジャガイド: 統計プロシジャ: CORR プロシジャ”, https://www.sas.com/offices/asiapacific/japan/service/help/webdoc/procstat/viewer.htm#corr_toc.htm, 2012, Accessed Aug 16, 2023

[13] SAS Institute, “SAS Blogs: How to interpret a residual-fit spread plot”, https://blogs.sas.com/content/iml/2013/06/12/interpret-residual-fit-spread-plot.html, 2013, Accessed Aug 16, 2023

[14] Social Survey Research Information Co., Ltd., “重回帰分析-エクセル統計による解析事例｜ブログ｜統計WEB”, 2017, Accessed Aug 15, 2023

治療群の選択を伴うアダプティブデザインの動作特性の検討

－事例に基づくシミュレーションの実践－

○高津　正寛[1,6]，飯塚　政人[2,6]，棚瀬　貴紀[3,6]，中村　将俊[4,6]，菅波　秀規[5,6]

（[1]持田製薬株式会社，[2]田辺三菱製薬株式会社，[3]大鵬薬品工業株式会社，[4]ファイザーR&D合同会社，[5]興和株式会社，[6]日本製薬工業協会 医薬品評価委員会 データサイエンス部会）

Evaluation of Operational Characteristics of Adaptive Designs with Treatment Selection

[1,6]Masahiro Takatsu, [2,6]Masato Iizuka, [3,6]Takanori Tanase, [4,6]Masatoshi Nakamura, [5,6]Hideki Suganami

[1]Mochida Pharmaceutical Co., Ltd., [2]Mitsubishi Tanabe Pharma Corporation, [3]Taiho Pharmaceutical Co., Ltd., [4]Pfizer R&D Japan G.K., [5]Kowa Company, Ltd., [6]Data Science Expert Committee, Drug Evaluation Committee, Japan Pharmaceutical Manufacturers Association

要旨

アダプティブデザインの一つに，治療群の選択に対するアダプテーションがある．複数の治療群を設定して試験を開始し，中間解析において事前に計画した選択基準に基づき治療群を選択し，最終解析において選択された治療群の検証を行うデザインであり，シームレス第 II/III 相デザインとも呼ばれる．

筆者らはアダプティブデザインに関する FDA ガイダンスと Mayer et al. (Stat Biopharm Res 2019, 11, 4, 325‐335)を参考に，シームレス第 II/III 相デザインの動作特性を評価するための指標を整理するとともに，組織内や規制当局にデザインの適切性を説明するために有用と考えられるシミュレーション報告書の構成を検討している．本稿では，シームレス第 II/III 相デザインの事例に基づき設定した複数の臨床シナリオに基づき，SAS のシミュレーションにより試験の動作特性を検討した結果を報告する．

キーワード：アダプティブデザイン，中間解析，シームレス第 II/III 相デザイン，シミュレーション，治療群選択

1　はじめに

ICH-E20 のトピックとして「アダプティブ臨床試験」が採択され，ガイドラインの作成が検討されている．2023 年 8 月時点において，European Medicines Agency と Food and Drug Administration（FDA）からアダプティブデザインのガイダンスが発行されている[5][6][7]．FDA ガイダンスでは，アダプティブデザインを「臨床試験に参加した被験者の蓄積されたデータに基づいて，試験デザインの 1 つ以上の側面について，予め計画された変更を行うことができる臨床試験デザイン」と定義している．アダプティブデザインにより与えられる柔軟性により，試験の参加者がより良い治療を受けられる機会が増え，より効率的な医薬品開発，さらには利用可能なリソースの活用といった利点が期待される．その一方，統計手法を適切に用いなければ，第一種

の過誤確率の増大，点推定値へのバイアスの発生，信頼区間の被覆確率が名義水準と異なるなど，統計的妥当性の観点から望ましくない現象が起こる可能性がある．日本製薬工業協会医薬品評価委員会データサイエンス部会は，アダプティブデザインの適切な実施を促進するために，FDA から公表されているアダプティブデザインに関するガイダンスの邦訳[18]と，アダプティブデザインの基本的な統計的推測法に関してまとめた「アダプティブデザインの統計的推測に関する検討」[19]を公表している．

FDA はさらに近年，「Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products」[8]というガイダンスを公表し，革新的な試験デザイン（CID）の取り組みを進めている．FDA は，CID に固定の定義は存在しないとしているが，事例の一つとしてアダプティブデザインを挙げている．また，「多くの CID に共通する一つの特徴は，試験の動作特性を推測するためには数式よりもシミュレーションが必要となることである 」と述べていることから，アダプティブデザインの動作特性を理解するためにはシミュレーションが重要であることは規制当局も認識していると考えられる．

アダプティブデザインの一つに，治療群の選択に対するアダプテーションがある．これは，複数の治療群を設定して試験を開始し，中間解析において事前に計画した選択基準に基づき治療群を選択し，最終解析において選択された治療群の検証を行うデザインであり，シームレス第 II/III 相デザインとも呼ばれる．

筆者らはアダプティブデザインに関する FDA ガイダンスと Mayer et al. (2019)[14]を参考に，治療群の選択を伴うシームレス第 II/III 相デザインや症例数の変更を伴うアダプティブデザインの動作特性を評価するための指標を整理するとともに，組織内や規制当局にデザインの適切性を説明するために有用と考えられるシミュレーション報告書の構成を検討している．本稿では，実際のシームレス第 II/III 相デザインに基づいた臨床シナリオを複数設定し，SAS のシミュレーションにより動作特性を検討した結果を報告する．

2　アダプティブデザインで用いられる解析手法

本項では，アダプティブデザインのうちシームレス第 II/III 相デザインにおいて用いられる解析手法である逆正規法と Step-Down Dunnett 検定について，手法の概要および SAS による実装方法を紹介する．

2.1　逆正規法

逆正規法は，各ステージから得られる独立した p 値を結合して検定を行う手法である[2][13]．2 ステージデザインの場合において，逆正規法の検定統計量Z_{INV}は以下の式で表され，標準正規分布に従う．

$$Z_{INV} = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)$$

ここで，$\Phi^{-1}(\cdot)$は標準正規分布の分布関数の逆関数，p_1およびp_2は各ステージのデータから得られる p 値である．w_1およびw_2は 2 つの各ステージの重みであり，$w_1^2 + w_2^2 = 1$を満たす範囲で自由に設定できる．各ステージで等しい重みとする$w_1 = w_2 = \frac{1}{\sqrt{2}}$という設定や，群ごとの各ステージの症例数$n_1$および$n_2$を重みに用いる$w_1 = \frac{n_1}{\sqrt{n}}, w_2 = \frac{n_2}{\sqrt{n}}$ $(n = n_1 + n_2)$という設定が考えられる．逆正規法に基づく最終解析の p 値は$p_{INV} = 1 - \Phi(Z_{INV})$で得られ，この p 値を有意水準と比較して検定を行う．正規分布に従うエンドポイントだけでなく，二値や生存時間などあらゆるエンドポイントに対して適用可能な手法である．治療群の選択を伴うデザインでは，複数の治療群とプラセボ群の比較のように多対一の比較を行う状況が想定され，Bonferroni 法や Dunnett 検定などにより適切に多重性を調整したステージごとの p 値を逆正規法の導出に用いる必要がある．

SAS では，データステップにおいて標準正規分布の分位点を返す関数 probit およびその逆関数 probnorm を利用して逆正規法の p 値を導出できる．R の rpact パッケージは，逆正規法による解析を実装可能である[16].

2.2 Step-down Dunnett 検定

Step-down Dunnett 検定は，多対一の比較において用いられるシングルステップの Dunnett 検定[3]を閉検定手順によって改良した手法である[4] [17]．Dunnett 検定と同様に，連続量のエンドポイントに対してのみ適用可能な手法である．

例として，試験治療群 3 用量（低，中，高用量群）とプラセボ群の 4 群の固定デザイン（治療群選択なし）を考える．各群の評価項目の平均値をそれぞれ$\mu_L, \mu_M, \mu_H, \mu_P$とする．3 つの基本帰無仮説（elementary null hypotheses）をそれぞれH_L, H_M, H_Hとし，$H_L : \mu_L = \mu_P, H_M : \mu_M = \mu_P, H_H : \mu_H = \mu_P$である．さらに，複数の帰無仮説の共通部分である積帰無仮説（intersection null hypothesis）として$H_{LM}, H_{LH}, H_{MH}, H_{LMH}$を考える（$H_{LM} : \mu_L = \mu_M = \mu_P$，$H_{LH} : \mu_L = \mu_H = \mu_P, H_{MH} : \mu_M = \mu_H = \mu_P, H_{LMH} : \mu_L = \mu_M = \mu_H = \mu_P$）．ここで，例えば低用量群のプラセボ群に対する優越性に興味があるとして，基本帰無仮説H_Lを棄却するためには，H_Lが属する全ての積帰無仮説（H_{LM}, H_{LH}, H_{LMH}）を棄却する必要がある．最初に，H_{LMH}について 3:1 の Dunnett 検定を有意水準 α にて行う．3 用量のいずれも有意でなかった場合，H_{LMH}を棄却できずに手順を終了する．3 用量のいずれかが有意であった場合，ひとつ下の積仮説H_{LM}およびH_{LH}について，2:1 の Dunnett 検定を有意水準 α にて行う．H_{LM}とH_{LH}のいずれも棄却された場合，基本帰無仮説H_Lについて，t 検定を有意水準 α にて行う．

Step-down Dunnett 検定は固定デザインの統計手法であるが，治療群の選択を伴うアダプティブデザインにおいては，第 1 ステージで選択されなかった群と対照群の差の検定統計量を$-\infty$と置き換えることで適用可能である[10][11]. これにより, 実際に第 2 ステージに進んだ群の数よりも多い数での対比を行うこととなり，検定の棄却限界値が大きくなるため，ファミリーワイズエラーを名義水準 α 以下に保つことが出来る．そのため，中間解析にて脱落した治療群が多いほど，保守的な結果を与える統計手法である．ただし，この手法では治療群選択以外のアダプテーションを適用することは出来ない．

SAS では，orthoreg プロシジャを利用して Step-down Dunnett 検定による解析が可能である．治療群の選択を伴うアダプティブデザインでは，あらかじめデータステップにて選択されなかった治療群の第 1 ステージにおけるエンドポイント変数に充分大きな数を減算（小さいほど改善を意味するエンドポイントの場合は，加算）しておくことで適用可能である．

```
proc orthoreg;
  class arm;
  model res=arm;
  lsmeans arm/pdiff=controlu("0") adjust=dunnett stepdown alpha=0.025;
run;
```

R では，multcomp パッケージを用いて Step-down Dunnett 検定が実行可能である[1]．治療群の選択を伴うアダプティブデザインへ適用するためには，SAS と同様にデータの前処理が必要となる．

3 アダプティブデザインの事例

治療群の選択を伴うアダプティブデザインを適用した事例として，慢性閉塞性肺疾患の患者を対象とした indacaterol の二重盲検ランダム化比較試験（INHANCE 試験）がある[12] [18]．INHANCE 試験は，慢性閉塞性肺疾患の患者を対象に，indacaterol 4 用量（75, 150, 300, 600μg o.d.）を試験治療，プラセボと formoterol および tiotropium を対照治療とした試験であり，2 ステージのアダプティブデザインを採用している．第 1 ステー

ジの目的は，第 2 ステージにて検証する indacaterol の用量を選択することであり，試験全体の主目的は第 1 ステージにて選択された少なくとも 1 つの indacaterol の用量のプラセボに対する優越性の検証，重要な副次目的は少なくとも 1 つの indacaterol の用量の tiotropium に対する非劣性の検証であった．中間解析における用量選択の基準として以下の 2 つを設定していた．

- 基準 1 : 2 週時点のトラフ FEV_1 を評価項目として，プラセボとの群間差の点推定値が 0.12 L よりも大きく，かつ点推定値が formoterol 群および tiotropium 群のいずれよりも大きい
- 基準 2 : 2 週時点の $FEV_1AUC_{1\text{-}4h}$ を評価項目として，点推定値が formoterol 群および tiotropium 群のいずれよりも大きい

2 つの評価項目トラフ FEV_1 と $FEV_1AUC_{1\text{-}4h}$ はいずれも数値が大きいほど効果が高いことを意味する．2 つの基準を満たす用量が複数ある場合，両基準を満たす最低の用量とそれより 1 つ高い用量が第 2 ステージに進むなど，2 つの基準を満たす状況に応じた用量選択ルールを事前に規定していた．なお，中間解析の目的は用量選択のみであり，症例数再推定等は計画していなかった．中間解析は 770 名（各群約 110 名）が 2 週間の治療を終えたタイミングで実施すると計画されていた．

その結果，中間解析に基づき選択された indacaterol 2 用量（150μg, 300μg），プラセボおよび tiotropium の計 4 群が有効性，安全性および忍容性を検討する 26 週間の第 2 ステージに進むとともに，追加の被験者組み入れが行われた．有効性の最終解析には 12 週時点のトラフ FEV_1 が評価項目に用いられ，各群 376～393 例が解析対象となった．試験開始時の indacaterol が 4 用量であったため，Bonferroni 法に基づき有意水準を $\alpha/4$（片側 $\alpha=0.025$）として indacaterol 2 用量のプラセボに対する優越性の検定が行われた．より検出力の高い方法を適用することは可能であったが，試験計画が複雑になるため採用しなかったとしている．最終解析の結果，indacaterol 2 用量ともにプラセボに対する優越性が検証された．

4　臨床シナリオおよび評価指標の設定

本項では，INHANCE 試験を下地に検討した臨床シナリオを説明する．想定する臨床シナリオは，オリジナルの INHANCE 試験から主に 3 つの点を変更している．第一に，最終解析の目的は indacaterol のプラセボ群に対する優越性検証のみとし，tiotropium 群に対する非劣性検証は検討対象外とした．第二に，陽性対照群は両ステージを通して 1 群のみとし，indacaterol の用量選択の判定にのみ用いた．第三に，用量選択と最終解析に用いる評価項目はいずれも 12 週時点のトラフ FEV_1 とし，2 週時点のトラフ FEV_1 および $FEV_1AUC_{1\text{-}4h}$ は検討対象外とした．そのため，オリジナルと異なり用量選択と最終解析に同一のデータを使用している．

用量選択ルールとして，オリジナルの INHANCE 試験に基づき，有効性の基準を満たした最低用量とそれより 1 つ高い用量を選択するルールを設定した．なお，1 つ高い用量については有効性の基準を満たしていなくとも選択可能である．加えて，INHANCE 試験終了後の FDA との審査時のやり取りにおいて，選択されなかった 75μg と選択された 2 用量（150μg, 300μg）とで有効性の差異が小さかったため，75μg について追加の検討がなされるべきであった[9]と審査報告書に記載があった点を考慮し，前述のルールに加えて，有効性の基準を満たした最低用量とそれより 1 つ低い用量を選択するルールも設定した．

シミュレーションの入力設定

パラメータ	設定
有意水準 (α)	0.025 (片側)
群	第 1 ステージ：6 群（試験治療 4 用量，プラセボ群，陽性対照群） 第 2 ステージ：3～4 群（選択された試験治療 1～2 用量，プラセボ群，陽性対照群）
第 1 ステージの例数/群	110
第 2 ステージの例数/群	230
割付比	全ての群で同じ
中間解析のタイミング	第 1 ステージ終了時
中間解析の目的	最終解析において検証する試験治療の選択 ※有効中止を意図した中間解析の実施はなし
用量選択ルール	選択条件： 「FEV1 のプラセボ群との差がエフェクトサイズとして 0.4 よりも大きい」かつ「陽性対照群との差が 0 より大きい」 選択ルール 1 : 選択条件を満たす最低用量①と，①より 1 つ高い用量②の **2 群** 選択ルール 2 : 選択条件を満たす最低用量①と，①より 1 つ低い用量②の **2 群** ※②については選択条件を満たしていなくても選択可能とし，②がなければ 1 用量のみ選択する
最終解析の目的	中間解析において選択された試験治療のプラセボ群に対する優越性の検証
検討する試験デザイン案および統計手法	1 Inferentially Seamless：最終解析に第 1 ステージおよび第 2 ステージのすべてのデータを使用する，治療群選択のアダプティブデザイン 1.1 逆正規法 ※重みには群ごとの各ステージの症例数を用いる．ステージごとの多重調整には Bonferroni 法を用いる 1.2 Step-down Dunnett 検定 2 Operationally Seamless：最終解析には第 2 ステージのデータのみ使用する，治療群選択の固定デザイン（ベンチマークとして設定） 2.1 Step-down Dunnett 検定 いずれのデザインも，推定手法はナイーブな方法（未調整の平均値）を用いる．

なお，Inferentially Seamless デザインと Operationally Seamless デザインの両方において Step-down Dunnett 検定を用いるが，結果の叙述では Inferentially Seamless デザインの方のみ Step-down Dunnett 検定と表記し，Operationally Seamless デザインについては検定名の表記を省略する．

第一種の過誤確率評価用のシナリオと検出力評価用のシナリオをそれぞれ設定した．$\mu_1, \mu_2, \mu_3, \mu_4$は用量 1～4 の，$\mu_9$は陽性対照群の治療効果（エフェクトサイズ）を表している．シミュレーション回数は，FDA のアダプティブデザインのガイダンスに「シナリオごとに繰り返し回数を 100,000 回とすると，第一種の過誤

確率の両側 95%信頼区間の幅が約±0.1%になることが保証され，ほとんどの場合は十分な回数となる」と記載されている[7][18]ことを踏まえ，100,000 回とした．なお，紙面の都合上，検出力評価用のシナリオは 201～204 の結果のみを本稿に示す．205～208 の結果は発表資料を参照のこと．

<u>第一種の過誤確率評価用のシナリオ</u>

No.	$(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\mu_3}, \boldsymbol{\mu_4}, \boldsymbol{\mu_9})$	備考
100	(0, 0, 0, 0, 0)	全用量効果なし
101~104	(**0.5**, 0, 0, 0, 0),⋯, (0, 0, 0, **0.5**, 0)	1 用量効果あり
105~110	(**0.5**, **0.5**, 0, 0, 0),⋯, (0, 0, **0.5**, **0.5**, 0)	2 用量効果あり
111~114	(**0.5**, **0.5**, **0.5**, 0, 0),⋯, (0, **0.5**, **0.5**, **0.5**, 0)	3 用量効果あり

<u>検出力評価用のシナリオ</u>

No.	$(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\mu_3}, \boldsymbol{\mu_4}, \boldsymbol{\mu_9})$	備考
201	(0.4, 0.5, 0.5, 0.5, 0.4)	用量 2 で頭打ち
202	(0.4, 0.45, 0.5, 0.5, 0.4)	用量 3 で頭打ち
203	(0.4, 0.433, 0.467, 0.5, 0.4)	用量反応
204	(0.4, 0.4, 0.5, 0.5, 0.4)	用量 1＝ 用量 2
205	(0.4, 0.5, 0.5, 0.5, 0.3)	201-204 より陽性対照群の効果が 0.1 低い
206	(0.4, 0.45, 0.5, 0.5, 0.3)	
207	(0.4, 0.433, 0.467, 0.5, 0.3)	
208	(0.4, 0.4, 0.5, 0.5, 0.3)	

検討した動作特性の評価指標は以下の通り．

<u>動作特性の評価指標</u>

指標	目的
1. 第一種の過誤確率	第一種の過誤確率評価用のシナリオの全てにおいて，検討する試験デザイン案/統計手法案における第一種の過誤確率（効果なしのいずれかの用量が選択されかつプラセボ群との有意差が得られる確率）が名義水準以下になることを確認する．
2. 検出力	検出力評価用のシナリオにおいて，検討する試験デザイン案/統計手法案における検出力を確認する． ・用量ごとの検出力 ・用量群全体の検出力（効果ありのいずれかの用量が選択されかつプラセボ群との有意差が得られる確率）
3. 各群と試験全体の症例数の分布	検出力評価用のシナリオにおいて，各群および試験全体における症例数の分布を確認する．
4. 各群の選択割合，選択用量数の分布	検出力評価用のシナリオにおいて，各群が選択される割合および選択用量数の分布を確認する．設定した用量選択ルールが期待通りに機能していることを確認する．選択用量数が 0 である割合は，基準を満たす用量が存在せずに試験を中止する無効中止確率に相当する．
5. 治療効果の推定	検出力評価用のシナリオにおいて，ステージごとおよび両ステージにおける，各群の点推定値（平均値）のバイアス，MSE および信頼区間の被覆確率を評価する． 中間解析にて選択された場合および選択されなかった場合における，条件付き推定の結果の評価も行う．

5　結果と考察

1.　第一種の過誤確率

用量選択ルールおよび試験デザイン/統計手法ごとの全体の第一種の過誤確率を表 1 に示した．第一種の過誤確率（全体）は，シナリオ No.100~114 の全シナリオにおける第一種の過誤確率（効果なしのいずれかの用量が選択されかつプラセボ群との有意差が得られる確率）のうちの最大値を意味している．いずれの用量選択ルールおよび統計手法においても，第一種の過誤確率が強い意味で制御されていた．また，Operationally Seamless デザインよりも Inferentially Seamless デザインの 2 手法がより保守的となり，Step-down Dunnett 検定よりも逆正規法（Bonferroni 法による調整）の方が保守的であることが確認された．

表 1 第一種の過誤確率（全体）

試験デザイン/統計手法	用量選択ルール	
	高い用量を選択	低い用量を選択
逆正規法	0.32%	0.34%
Step-Down Dunnett 検定	0.83%	0.80%
Operationally Seamless	1.66%	1.67%

2.　検出力

各用量選択ルールにおける検出力をそれぞれ図 1 および図 2 に示した（数値は発表資料を参照）.

高い用量を選択するルールにおける用量群全体の検出力（効果ありのいずれかの用量が選択されかつプラセボ群との有意差が得られる確率）について，シナリオ 201～204 の全てで検出力が 80%を超えていた．手法間の差異は微々たるものであったが，いずれのシナリオでも Operationally Seamless デザインより Inferentially Seamless の 2 手法においてわずかに検出力が高い様子が見られた．

低い用量を選択するルールにおいても，用量群全体の検出力では同様の傾向が見られた．群ごとの検出力では，用量選択ルールを反映して低い用量における検出力が高くなる傾向が見られた．

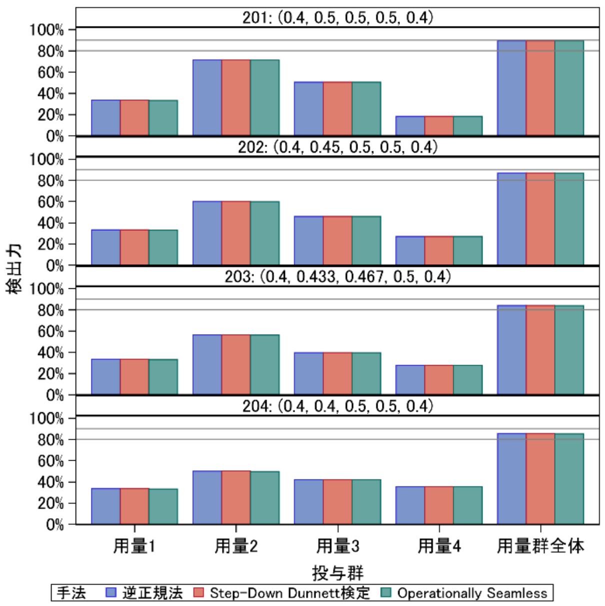


図 1 検出力（高い用量を選択）

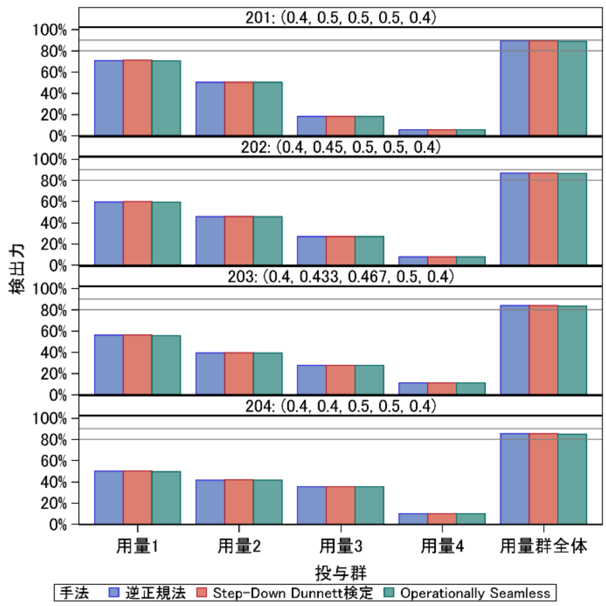


図 2 検出力（低い用量を選択）

3.　各群と試験全体の症例数の分布

各用量選択ルールにおける各群および 6 群全体の両ステージにおける症例数の分布を評価した（図および数値は発表資料を参照）．なお，いずれの用量選択ルールでも，症例数の最小値は第 1 ステージにおいて 1 用量も選択されなかった場合（110 例×6 群）であり，最大値は 2 用量ともに選択された場合（110 例×6 群+230 例×4 群）となる．

高い用量を選択するルールにおける選択用量数について，シナリオ 201 にて最も期待症例数（症例数の平均値）が多く（1469.0 例），シナリオ 203 にて最も期待症例数が少なかった（1407.0 例）．低い用量を選択するルールでは，シナリオ 201 にて最も期待症例数が多く（1404.9 例），シナリオ 203 にて最も期待症例数が少なかった（1355.7 例）．いずれのシナリオにおいても高い用量を選択するルールと比較して低い用量を選択するルールにて期待症例数が少なくなっており、選択用量ルールを反映している様子が確認された．

4.　各群の選択割合，選択用量数の分布

各用量選択ルールにおける各群の選択割合をそれぞれ図 3 および図 4 に，各用量選択ルールにおける選択用量数の分布をそれぞれ図 5 および図 6 に示した．

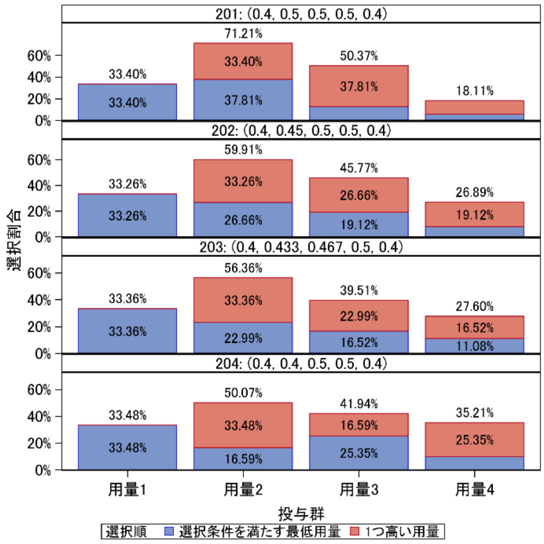


図 3 各群の選択割合（高い用量を選択）

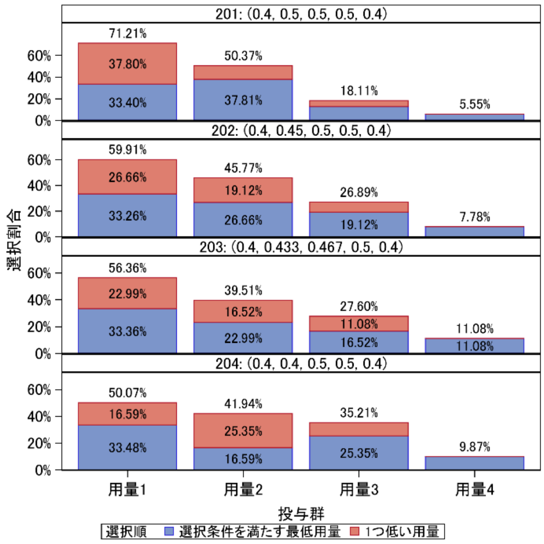


図 4 各群の選択割合（低い用量を選択）

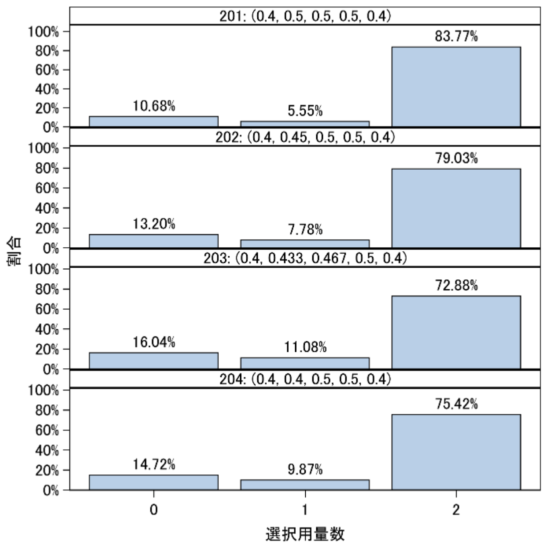


図 5 選択用量数（高い用量を選択）

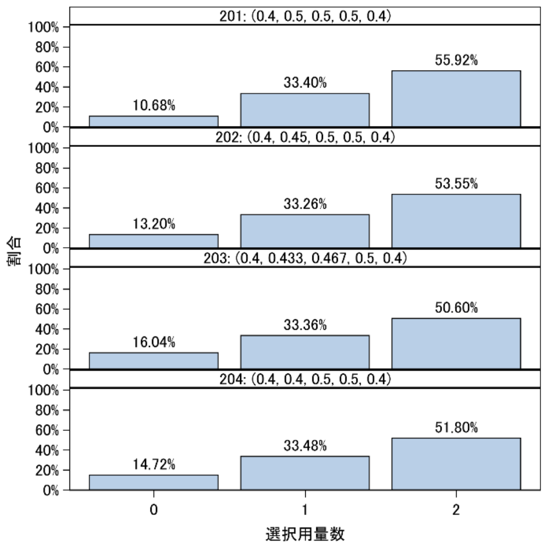


図 6 選択用量数（低い用量を選択）

高い用量を選択するルールにおいては，最低用量の用量 1 はすべて「選択条件を満たす最低用量」としてのみ選択されている一方，用量 2 から 4 は「選択条件を満たす最低用量」として選択される場合と，「その 1 つ高い用量」として選択される場合の 2 通りがある．いずれのシナリオにおいても，用量 2 の選択割合が一番高いことが分かる．低い用量を選択するルールにおいても同様に，最高用量の用量 4 はすべて「選択条件を満たす最低用量」としてのみ選択されている一方，用量 1 から 3 は「選択条件を満たす最低用量」として選択される場合と，「その 1 つ低い用量」として選択される場合の 2 通りがある．いずれのシナリオにおいても，用量 1 の選択割合が一番高いことが分かる．なお，用量選択ルールの定義から，「選択条件を満たす最低用量」として選択される割合は両ルールで同一となっており，高い用量を選択するルールにおける用量 2~4 の選択割合は，低い用量を選択するルールにおける用量 1~3 の選択割合とそれぞれ同一となっている．

選択用量数の分布に関して，高い用量を選択するルールにおいては 7 割以上の確率で 2 用量が選択されている一方，低い用量を選択するルールにおいては 5 割前後にとどまっている．なお，前述のとおり，「選択条件を満たす最低用量」として選択される割合が両ルールで同一のため，無効中止（選択用量数 0）が発生する割合も両ルールで同一となっている．

5.　<u>治療効果の推定</u>

各用量選択ルールにおけるステージごとおよび両ステージの推定値のバイアス（条件なし）をそれぞれ図 7 および図 9 に，第 2 ステージに選択された条件付き，もしくは選択されなかった条件付き（第 1 ステージのみ）のバイアスをそれぞれ図 8 および図 10 に示した．MSE および被覆確率については発表資料を参照．なお，プラセボ群および陽性対照群における「選択されなかった条件付き」とは，1 用量も選択されず無効中止となった条件付きを意味している．また，第 2 ステージについてはおのずと選択された条件付きとなるため，「条件なし」と「選択された条件付き」とで同一の結果を示している．

高い用量を選択するルールにおいて，条件なしの場合（図 7），第 1 ステージおよび第 2 ステージそれぞれの推定値のバイアスは 0 に近い値となっていた．一方，両ステージの推定値においてプラセボ群および陽性対照群では正のバイアス，各用量群では負のバイアスがわずかに含まれていた．ここで，条件なしの両ステージにおける推定値には，当該群が選択された場合は両ステージの，選択されなかった場合は第 1 ステージの推定値を使用しているため，それぞれの条件付きに分けた結果を確認することにする．

選択された条件付きの場合（図 8），第 1 ステージの各用量群の推定値に正のバイアス，プラセボ群および陽性対照群ではわずかな負のバイアスが含まれており，その影響で両ステージの推定値にもバイアスが含まれている様子が見られた．これは，選択条件（プラセボ群との差が 0.4 よりも大きいかつ陽性対照群との差が 0 より大きい）を満たすのは当該用量群の効果が真値よりも高く，プラセボ群の効果が真値よりも低いためである．用量ごとのバイアスに注目すると，最低用量の用量 1 におけるバイアスが他の用量群と比較して大きいが，これは用量 1 が選択されるケースは必ず選択条件を満たす最低用量として選択されるため，低い効果が得られた場合は選択されない一方，用量 2~4 については選択条件を満たす最低用量として選択されるケースと，それより 1 つ高い用量として選択されるケースとが混在しているためである．選択されなかった条件付きの場合（図 8）は，選択された条件付きと逆の理由から，第 1 ステージの各用量群の推定値に負のバイアス，プラセボ群および陽性対照群では正のバイアスが含まれていた．条件なしの場合（図 7）における両ステージの推定値のバイアスにはこれらが反映されたものと考えられた．すなわち，各用量群においては，選択された条件付きにおいて得られた両ステージの正のバイアスよりも，選択されなかった条件付きにおいて得られた第 1 ステージの負のバイアスの方が絶対値が大きいために負のバイアスとなり，プラセボ群および陽性対照群はその逆となったと解釈できる．例えば，シナリオ 201 における用量 1 は 100,000 回のシミュレ

ーションのうち 33,401 回選択され，選択された条件付きの両ステージのバイアスは 0.026，選択されなかった条件付きの第 1 ステージのバイアスは-0.041 から，条件なしの両ステージのバイアスはこれらの重み付き平均として，$(33{,}401 \times 0.026 + 66{,}599 \times (-0.041))/100{,}000 \fallingdotseq -0.018$となる．

低い用量を選択するルールにおける選択用量数について，条件なしの場合（図 9）および条件付きの場合（図 10）の傾向は，高い用量を選択するルールの場合と同様だった．選択されたという条件付きの用量ごとのバイアスに注目すると，最高用量の用量 4 におけるバイアスが大きく，高い用量を選択するルールと同様に，用量 4 が選択されるケースは必ず選択条件を満たす最低用量として選択され，低い効果が得られた場合は選択されないためと考えられた．

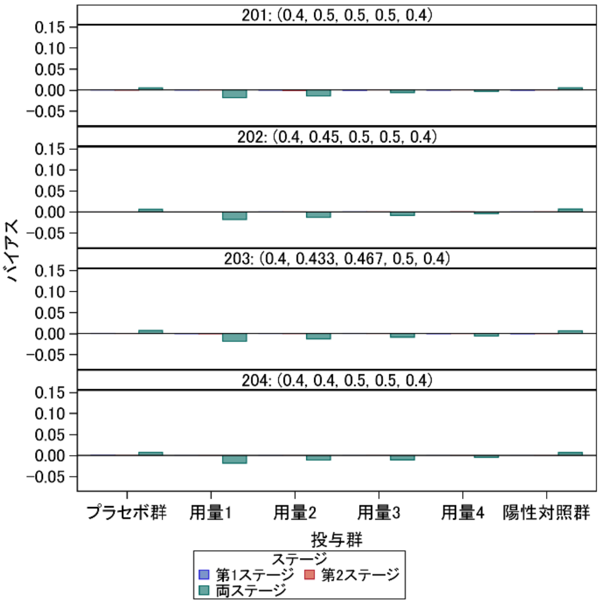


図 7 バイアス（高い用量を選択）

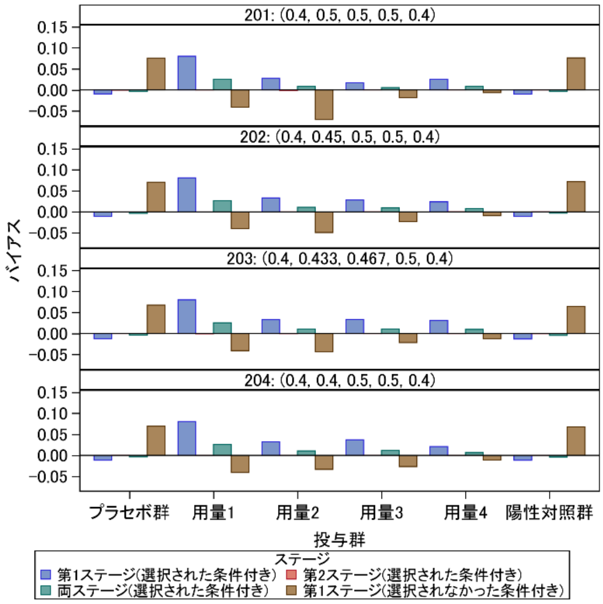


図 8 バイアス（条件付き；高い用量を選択）

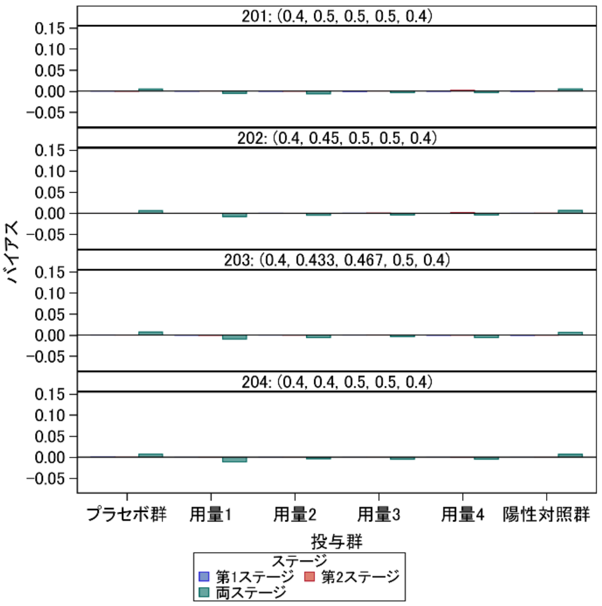


図 9 バイアス（低い用量を選択）

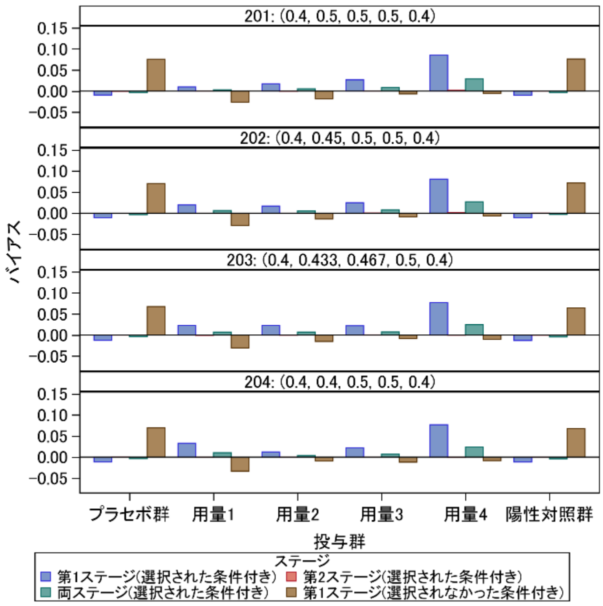


図 10 バイアス（条件付き；低い用量を選択）

6　まとめ

本稿では，治療群の選択を伴うアダプティブデザインの事例に基づき設定した臨床シナリオに基づき，動作特性を評価した．第一種の過誤確率においてはいずれのデザインも名義水準を下回っていたが，Operationally Seamless デザインの方が名義水準に近い値となった．検出力の検討においてはデザインおよび統計手法間の差異は微々たるものであった．治療群の選択ルールについては，薬剤特性などの臨床的な観点も踏まえて適切な手法を選択すべきであるが，シミュレーションにより動作特性を明らかにし，統計学的な側面も踏まえて検討することは選択ルールの設定において重要と考えられる．あわせて，検出力の評価においては用量に応じた効果が単調増加（非減少）するシナリオのみ検討したが，途中の用量における効果が一番大きく，それより高い用量では効果が下がるという二次関数的な用量反応曲線なども考えられ，疾患や薬剤の特性によって期待される用量反応関係を幅広く検討する必要がある．治療効果の推定においては，第 1 ステージで用量選択条件を満たしたという条件付きの推定値に過大評価のバイアスが確認されたため，推定値の解釈に留意するとともに，必要に応じてバイアスを補正する手法の利用も検討する．

実際にアダプティブデザインを利用するにあたっては，試験の完全性に与える影響や，試験結果の解釈を複雑にするリスク，運用上の煩雑さやコスト等も考慮したうえで総合的に検討する必要があると考えられる．そのため，検出力や症例数などの統計学的な側面のみに基づきアダプティブデザインの利用を検討することは推奨されない．しかし，幅広い臨床シナリオの下でシミュレーションに基づきアダプティブデザインの動作特性を明らかにし，シミュレーション報告書を構成することは，組織内や規制当局との議論を円滑化し，アダプティブデザインの導入の障壁を下げる一助となることが期待される．

参考文献

[1] Bretz F, Hothorn T, Westfall P. Multiple comparison using R. Chapman and Hall/CRC.

[2] Chow S-C and Chang M. Adaptive design methods in clinical trials. Chapman and Hall/CRC, 2006. 平川　晃弘，五所　正彦訳．臨床試験のためのアダプティブデザイン．朝倉書店．東京．2018.

[3] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 1955;50:1096–1121.

[4] Dunnett CW, Tamhane AC. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. Stat Med. 1991 Jun;10(6):939-47.

[5] European Medicines Agency. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design (CHMP/EWP/2459/02) 2007.

[6] Food and Drug Administration. Adaptive Designs for Medical Device Clinical Studies Guidance for Industry and Drug Administration Staff. 2016.

[7] Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. 2019.

[8] Food and Drug Administration. Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products Guidance for Industry. 2020.

[9] Food and Drug Administration. Medical Review, APPLICATION NUMBER:022383Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022383Orig1s000MedR.pdf [accessed 2023/8/28]

[10] Friede T, Stallard N. A comparison of methods for adaptive treatment selection. Biometrical Journal 2008;50(5):767-

81.
[11] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in Medicine 2008;10;27(10):1612-25.
[12] Lawrence D, Bretz F, Pocock S. INHANCE: An adaptive confirmatory study with dose selection at interim. In Indacaterol - The First Once-Daily Long-Acting Beta2 Agonist for COPD, Trifilieff A (ed.) Springer: Basel, 2014; 77–93.
[13] Lehmacher W, Wassmer G. Adaptive sample-size calculations in group sequential trials. Biometrics 1999;55:1286-1290.
[14] Mayer C, Perevozskaya I, Leonov S, Dragalin V, Pritchett Y, Bedding A, Hartford A, Fardipour P, Cicconetti G. Simulation Practices for Adaptive Trial Designs in Drug and Device Development. Statistics in Biopharmaceutical Research 2019;11(4):325-335.
[15] Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. Biometrics. 2001 Sep;57(3):886-91.
[16] Rpact package. https://www.rpact.com/vignettes/rpact_mams_design_and_analysis [accessed 2023/8/4]
[17] 土居 正明．多重性制御の基礎理論（閉検定手順）．計量生物学 Vol. 36, Special Issue, S 99–S 121 (2015).
[18] 日本製薬工業協会．アダプティブデザインに関する FDA ガイダンスの邦訳．2021 年 8 月. https://www.jpma.or.jp/information/evaluation/results/allotment/adaptive_design.html [accessed 2023/8/4]
[19] 日本製薬工業協会．アダプティブデザインの統計的推測に関する検討．2023 年 2 月. https://www.jpma.or.jp/information/evaluation/results/allotment/DS_202302_adaptive.html [accessed 2023/8/4]

防災情報は住民に的確に届いているか？
―全国ウェブ調査による実態把握とJMPによる分析－

○有馬昌宏，川向肇，阿部太郎

（兵庫県立大学）

Are Residents Recognize Their Natural Disaster Risk Correctly?

Masahiro Arima, Hajime Kawamukai and Taro Abe
University of Hyogo

要旨

近年，豪雨や台風による洪水や土砂災害などの大規模な自然災害が頻発している．このような状況のもと，国や自治体はハザードマップによる被災可能性の事前の把握と危険がある場合は早めの事前避難を推奨している．しかし，自然災害による被害は減少しておらず，そもそも，住民がハザードマップを読んで被災可能性の有無を的確に判断しているかどうかに疑問が生じている．そこで，2022 年 12 月から翌 1 月にかけて，全国ウェブ調査を実施した．この調査では，第１段階で自宅の自然災害に対する脆弱性の有無と有の場合はその具体的内容（想定浸水深など）を回答してもらい，第２段階でウェブ調査を中断して我々が開発を続けている防災アプリ「ハザードチェッカー」で自然災害に対する脆弱性を確認し，第３段階で再びウェブ調査に戻って，認識と実際の確認結果の間に齟齬がないかを問う質問を調査の中心に据えており，1,311 人からの回答を得ている．本発表では，このウェブ調査の概要を紹介し，認識と実際が異なることが判明した 15.3%の回答者に焦点を当て，その齟齬をもたらした原因について，災害種別に分けて実施した JMP を用いた要因分析の結果を報告する．

キーワード：防災，ハザードマップ，読図，誤認識，全国ウェブ調査，防災アプリ

１．はじめに

ソフト防災を有効に機能させるためには，個々の居宅や事業所などが立地する特定の地点を対象とした避難情報の発令が有用であると考えられる．しかし，地点別の避難情報の自治体からの発令は現状では難しい．

そこで，本発表では，特定の地点に特化した避難情報の提供に向けて，避難の判断に必要な地点特化の情報を高い情報品質で提供することを目指して 2015 年から開発を続けてきている防災アプリ「ハザードチェッカー」の現時点での改修内容を紹介するとともに，アプリの有用性と課題を評価するために実施した全国ウェブ調査の結果を報告し，「住民ひとりひとりに届ける危機対応避難情報提供アプリ」の実現に向けての取組

を紹介する．

２．情報品質の高いピンポイントの避難情報の必要性

防災対策は，堤防やダムや防潮堤などの構造物に依存するかどうかで大きく 2 つに分類される．1 つは堤防やダムや防潮堤や地盤の嵩上げなどの構造物に依存するハード防災であり，もう 1 つは構造物に依存せずに危険が迫った際の避難で対応するソフト防災である．

我々の研究チームは，ソフト防災が効果的に機能するには，

①所在地（特に居宅）に存在する素因（各種自然災害に対する脆弱性）の有無と素因が存在する場合の具体的な内容の正しい認識，

②存在する素因に対応する自然災害を惹起する誘因の接近・存在の迅速な把握，

③存在する素因と接近・発生している誘因の関係から自然災害の発生の可能性の正しい予測，

④状況に応じた在宅避難，垂直避難または水平避難（立ち退き避難）の選択判断と判断結果の迅速な実行，

が必要であると考えている．このようなソフト防災を効果的に機能するには，市町村から発令される避難情報も有用であるが，現状の避難情報は各地点の有する地理的特徴に応じてのピンポイントで有効な情報とはなっておらず，結果として避難情報に基づく避難率（実際の避難所への避難者数／避難情報の対象地域の居住者数）が低くなり，避難情報が効果的に利用されていない可能性が存在している．

このような問題意識に基づく研究として，避難情報の対象地域を絞り込むことの有効性に焦点を当てた廣井・保科(2020)などの研究がある．我々は，これらの研究で実証的に示される前から，究極の避難情報は個別の住居や事業所を対象にピンポイントで発令されるべきものであり，

ａ．広域ブロードバンドのインターネット網の基盤整備，

ｂ．スマートフォンなどの端末の普及，

ｃ．各災害リスクの種類別に Shape ファイルなどで提供されるデジタルのハザードに関する空間情報の整備とオープンデータとしての公開，

ｄ．気象庁の防災情報の XML フォーマット形式電文のリアルタイムでの公開，

という環境が整えば，個別の地点に特化した避難情報生成に繋がるデータ取得と住民や勤労・就学者へのピンポイントの防災（避難）に資するリスク情報の伝達を可能とする情報システムの構築が可能であると考えていた．

３．防災アプリ「ハザードチェッカー」の開発

我々の防災アプリ開発チームは，個別の地点に特化した防災関連情報をワンストップで提供する防災アプリのプロトタイプとして「ハザードチェッカー」の開発・運用を 2015 年に開始し，その後も提供する防災情報の情報品質の向上を目指して改修･改良に取り組んできている（田中･有馬(2016), 有馬(2017), 有馬他(2023a, 2023b)，川向他(2023)）．

しかし，開発開始から改修・改良を重ねながら 8 年が経過しようとしているところで，開発目的の情報品質の向上という視点から「ハザードチェッカー」の提供情報や情報表示方法や操作性を再検討してみたところ，本源性（内容が正しい），文脈性（判断に役に立つ），表現性（分かりやすい），利用性（利用しやすい）という情報品質を構成する 4 つの次元のうち，提供される情報が過多となり，操作の煩雑さの増加で表現性

表 1　防災アプリ「ハザードチェッカー」の一発確認画面で表示される情報と画面の遷移

		洪水	内水氾濫	高潮	津波	土砂災害	地震 揺れやすさ	地震 液状化	火災
素因ボタン	ハザード 表示内容	浸水想定区域 （計画規模・ 想定最大規模）	浸水想定区域 （広島市のみ）	浸水想定区域	浸水想定区域	土砂災害 警戒区域・ 危険箇所	地盤増幅率	微地形区分	密集市街地
	浸水深等 家屋表示	表示	表示	表示	表示	表示	なし	なし	なし
	使用データ	国土交通省 国土数値情報	広島市 下水道局	国土交通省 国土数値情報	国土交通省 国土数値情報	国土交通省 国土数値情報	J-SHIS　地震 ハザードカルテ	J-SHIS　地震 ハザードカルテ	国土交通省 国土数値情報
	点滅・振動 する条件	想定区域内	想定区域内	想定区域内	想定区域内	警戒区域・危険 箇所内	増幅率が2以上	液状化が起き やすい微地形	指定区域内
	タップの 遷移先 （指定地点 中心の 地図）	国土交通省 重ねるハザード マップの洪水 浸水想定区域	なし	国土交通省 重ねるハザード マップの高潮 浸水想定区域	国土交通省 重ねるハザード マップの津波 浸水想定区域	国土交通省 重ねるハザード マップの土砂 災害関連の 警戒区域等	防災科学技術 研究所 J-SHIS 断層マップ	なし	なし
誘因ボタン	対応する 注意報・ 警報等	洪水注意報 氾濫注意情報 洪水警報 氾濫警戒情報 氾濫危険情報 氾濫発生情報	大雨注意報 大雨警報に切り 換える可能性が 高い注意報 大雨警報 大雨特別警報	高潮注意報 高潮警報に切り 換える可能性が 高い注意報 高潮警報 高潮特別警報	津波注意報 津波警報 大津波警報	大雨注意報 大雨警報に切り 換える可能性が 高い注意報 大雨警報 大雨特別警報	なし	なし	乾燥注意報
	点滅・振動 する条件	上記情報発表	上記情報発表	上記情報発表	上記情報発表	上記情報発表	なし	なし	上記情報発表
	タップの 遷移先 （地図など）	気象庁 洪水キキクル	気象庁 浸水キキクル	当該市町村の 警報・注意報の ページ	当該市町村の 警報・注意報の ページ	気象庁 土砂キキクル	なし	なし	当該市町村の 警報・注意報の ページ
警戒レベルボタン	表示色	レベル2：黄 レベル3：赤 レベル4：紫 レベル5：黒	レベル2：黄 レベル3：赤 レベル5：黒	レベル2：黄 レベル3：赤 レベル4：紫	レベル3：赤 レベル4：紫	レベル2：黄 レベル3：赤 レベル4：紫 レベル5：黒	なし	なし	レベル2：黄
	タップの 遷移先 （地図）	国土地理院 洪水に対応の 指定緊急 避難場所	国土地理院 内水氾濫に 対応の指定 緊急避難場所	国土地理院 高潮に対応の 指定緊急 避難場所	国土地理院 津波に対応の 指定緊急 避難場所	国土地理院 崖崩れ・土石 流・地滑りに 対応の指定 緊急避難場所	国土地理院 地震に対応の 指定緊急 避難場所		国土地理院 大規模な火事に 対応の指定 緊急避難場所
家屋の図で示した浸水深の矢印をタップすると，それぞれ，指定地点を中心とする該当の災害の重ねるハザードマップの浸水想定区域図に別画面で遷移									

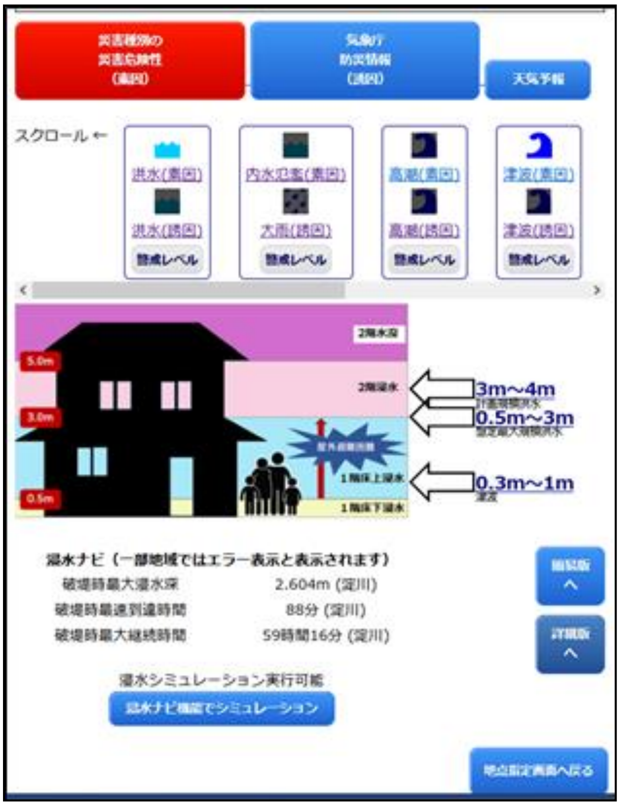


図 1　スマホの一発確認画面例（大阪駅を指定）

図 2　洪水ボタンタップでハザードマップ表示

図 3　PC の一発確認画面例（2023 年 8 月 15 日の台風 7 号接近時に神戸駅を指定）

が低下し，結果として避難の判断につながる文脈性も低下しているのではないかとの評価に至った．

そこで，2022 年 12 月に図 1 に示すような「一発確認画面」を導入して，GPS 機能がオンであれば，
①アプリのウェブサイト（http://urx3.nu/zk2F）にアクセス，
②「今ここ危険？」ボタンのタップ，
の 2 回の操作で，表 1 に示すように，指定地点に存在する素因と誘因の有無に関する情報と警戒レベルが当該ボタンの点滅と振動，黄・赤・紫・黒の色表示，浸水深表示用家屋図を用いた各種浸水被害の想定浸水深の矢印表示による取得ができ，各ボタンのタップで避難に役立つ情報が得られるよう改修を実施したところである．

また，警報レベル 3 以上の判定結果が得られた場合には，端末の振動や鳴動を可能とし，英語をはじめとする各種言語による表示など，利用者の視聴覚機能や言語にも配慮してのユニバーサル対応を図っているところである．

4．ウェブ調査の概要とリスク認識の齟齬の現状

自然災害に対する自宅のハザード（脆弱性）の有無，およびハザードが有りの場合の具体的な内容が正しく認識されているかどうかを確認するとともに，我々の提供している防災アプリ「ハザードチェッカー」の

機能の有効性評価を主たる目的として，全国ウェブ調査を実施した．ただし，ハザードの有無とハザードの具体的内容については，オープンデータとして公開されているハザードマップで提供されている情報に基づくという条件付きの調査である．調査の概要は以下の通りである．

調査名：防災アプリについてのアンケート

委託会社：株式会社データーサービス（西宮市）

調査対象：20 歳以上の日本国内在住者

実施期間：2022 年 12 月 28 日～2023 年 1 月 23 日

調査方法：調査用サイトを公開しての応募型

調査内容：3 つの段階で構成．

第 1 段階：調査用ウェブサイトで住所・性別・年齢の個人属性，構造や築年数などの居宅の属性，居住地選択理由，災害種別の災害リスクの有無の認識とその根拠，紙媒体のハザードマップの配布状況・閲覧経験・保管状況，防災アプリの利用状況，実施している防災対策などの質問に回答．

第 2 段階：「ハザードチェッカー」にアクセスして居宅の災害リスクを確認するとともに「ハザードチェッカー」を操作して各種機能を確認．

第 3 段階：再び調査用ウェブサイトに戻って，第 1 段階での居宅の災害リスクの認識と第 2 段階でのアプリでの災害リスクの有無の確認結果との間の齟齬とその内容，「ハザードチェッカー」の提供する 29 の機能の有用性の 5 段階評価，外出先での災害リスクの心配度，今後に予定の災害対策，同居家族数と同居家族属性，職業，被災経験の有無と被災による転居経験の有無，避難所避難経験の有無を回答．

回答者数：ウェブサイト訪問者 16,895 名，第 1 段階回答 2,220 名，第 3 段階回答 1,311 名，有効回答 1,270 名．

有効回答者の属性：

性別：男性（55.2%），女性（43.6%）．

年齢：20 代（4.5%），30 代（13.8%），40 代（21.6%），50 代（24.7%），60 代（22.3%），70 代以上（13.2%）．

住居構造：木造（54.2%），非木造（13.2%），鉄筋・鉄骨（29.9%）．

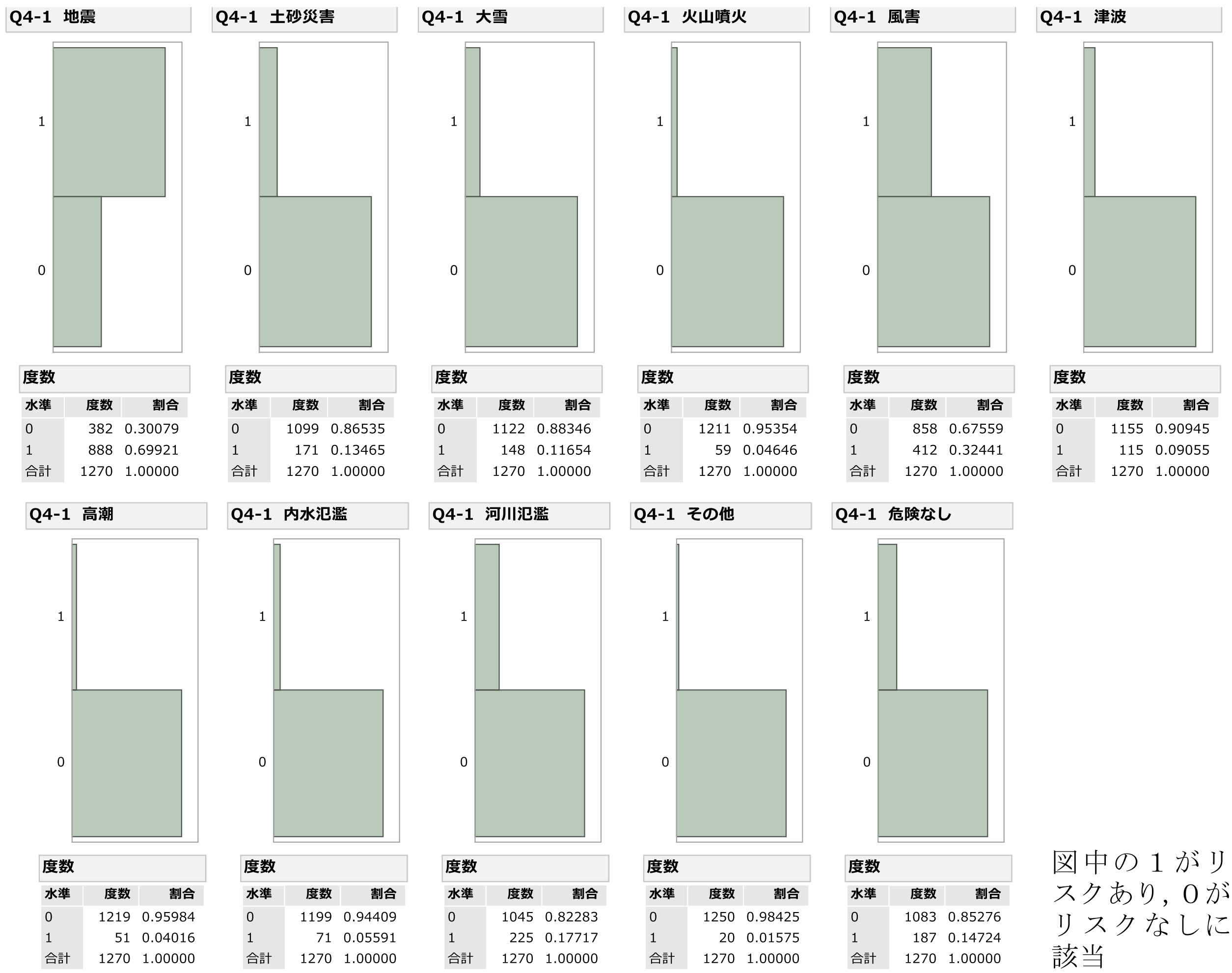
築年数：建築後 41 年以上の旧耐震基準（17.1%），建築後 41 年未満の新耐震基準（74.7%）．

以上の全国ウェブ調査では，面倒な 3 段階の手順を踏んで回答してもらうことにより，素因（自然災害の危険性）の有無と有りの場合のその内容について，「ハザードチェッカー」で危険性を確認する前と後での認識の齟齬を捉えることができるように設問している．

まず，特定の地点別に自然災害のリスクの有無とリスクが存在する場合の具体的なリスクの内容を簡単な操作で判定できる「ハザードチェッカー」で確認する前のリスク把握では，図 4 に示すように，自然災害に対するリスクがないと認識している回答者は 14.7%にとどまり，85.3%の回答者は何らかの自然災害のリスクを認識していることが示されている．また，認識されている自然災害の種別では，場所を選ばず，全国どこでも発生しうる地震および竜巻や台風による風害を認識している回答者の比率が高くなっている．

また，これらの自然災害のリスクの認識の根拠としては，図 5 に示すように，自分自身の経験や知識と自治体から配布される紙のハザードマップの回答比率が高く，本調査はウェブ調査で行われていることから回答者はスマートフォンや PC の操作に習熟している可能性が高いにもかかわらず，防災アプリによる認識の比率が 24.3%と，防災アプリを開発して提供している我々の想定よりも低いことが示されている．

ところで，既述のように，自宅の自然災害に対する脆弱性に関する認識が常に正しいとは限らない．ハザードマップを読図する際のミス，更新される前の古いハザードマップによる認識，長年の経験からの安心感などにより，認識されたリスクの有無およびリスクの内容と実際のリスクの有無およびリスクの内容には相



Q4-1 地震

水準	度数	割合
0	382	0.30079
1	888	0.69921
合計	1270	1.00000

Q4-1 土砂災害

水準	度数	割合
0	1099	0.86535
1	171	0.13465
合計	1270	1.00000

Q4-1 大雪

水準	度数	割合
0	1122	0.88346
1	148	0.11654
合計	1270	1.00000

Q4-1 火山噴火

水準	度数	割合
0	1211	0.95354
1	59	0.04646
合計	1270	1.00000

Q4-1 風害

水準	度数	割合
0	858	0.67559
1	412	0.32441
合計	1270	1.00000

Q4-1 津波

水準	度数	割合
0	1155	0.90945
1	115	0.09055
合計	1270	1.00000

Q4-1 高潮

水準	度数	割合
0	1219	0.95984
1	51	0.04016
合計	1270	1.00000

Q4-1 内水氾濫

水準	度数	割合
0	1199	0.94409
1	71	0.05591
合計	1270	1.00000

Q4-1 河川氾濫

水準	度数	割合
0	1045	0.82283
1	225	0.17717
合計	1270	1.00000

Q4-1 その他

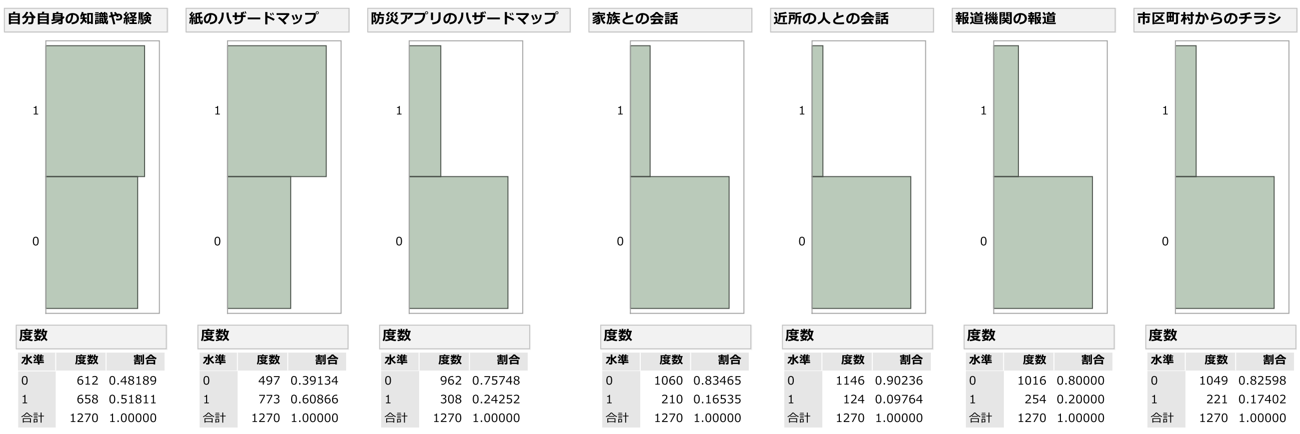
水準	度数	割合
0	1250	0.98425
1	20	0.01575
合計	1270	1.00000

Q4-1 危険なし

水準	度数	割合
0	1083	0.85276
1	187	0.14724
合計	1270	1.00000

図中の 1 がリスクあり，0 がリスクなしに該当

図 4　防災アプリで確認する前の自然災害対する脆弱性の認識



自分自身の知識や経験

水準	度数	割合
0	612	0.48189
1	658	0.51811
合計	1270	1.00000

紙のハザードマップ

水準	度数	割合
0	497	0.39134
1	773	0.60866
合計	1270	1.00000

防災アプリのハザードマップ

水準	度数	割合
0	962	0.75748
1	308	0.24252
合計	1270	1.00000

家族との会話

水準	度数	割合
0	1060	0.83465
1	210	0.16535
合計	1270	1.00000

近所の人との会話

水準	度数	割合
0	1146	0.90236
1	124	0.09764
合計	1270	1.00000

報道機関の報道

水準	度数	割合
0	1016	0.80000
1	254	0.20000
合計	1270	1.00000

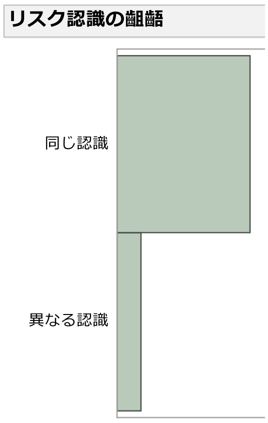
市区町村からのチラシ

水準	度数	割合
0	1049	0.82598
1	221	0.17402
合計	1270	1.00000

図 5　自宅の自然災害に対する脆弱性の認識の根拠

違が生じることがある．

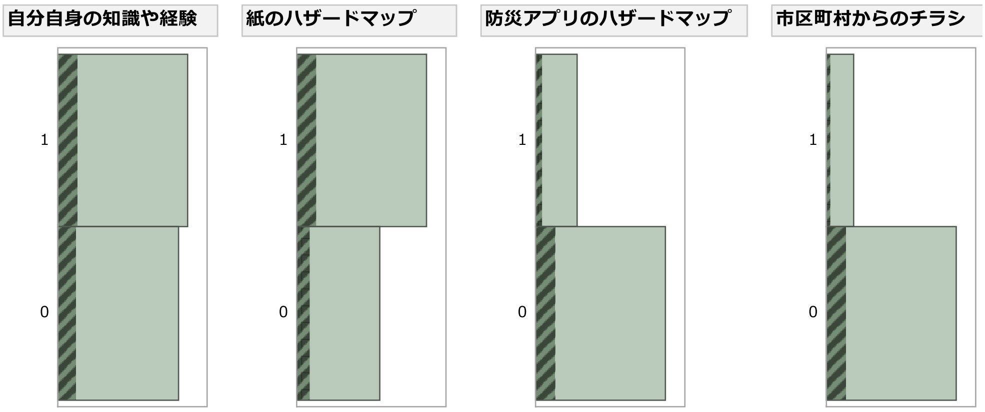
そのため，ウェブ調査への回答を一時中断し，「ハザードチェッカー」で自宅の自然災害に対する脆弱性の有無を確認するとともに，「ハザードチェッカー」の脆弱性判定結果画面に判定結果を 9 桁のコードで表示させ，そのコードの入力と合わせて，「ハザードチェッカー」の判定結果と自身の認識との間に齟齬があったかどうかを質問している．



度数

水準	度数	割合
異なる認識	194	0.15276
同じ認識	1076	0.84724
合計	1270	1.00000

図 6　認識の齟齬



度数（自分自身の知識や経験）

水準	度数	割合
0	612	0.48189
1	658	0.51811
合計	1270	1.00000

度数（紙のハザードマップ）

水準	度数	割合
0	497	0.39134
1	773	0.60866
合計	1270	1.00000

度数（防災アプリのハザードマップ）

水準	度数	割合
0	962	0.75748
1	308	0.24252
合計	1270	1.00000

度数（市区町村からのチラシ）

水準	度数	割合
0	1049	0.82598
1	221	0.17402
合計	1270	1.00000

図 7　リスク認識の根拠とリスク認識の齟齬との関係

その結果，図 6 に示すように，認識が異なったとの回答は 15.3%（認識が異なった災害種別の内訳は，複数回答で，洪水 9.7%，内水氾濫 6.1%，高潮 3.5%，津波 4.2%，土砂災害 4.9%，地震 7.0%，ただし災害種別によってはハザードマップが整備されていない市区町村（特に内水氾濫）やオープンデータとして公開されていない市区町村があるため，正確性に欠けることに留意が必要）を示しており，リスク認識の根拠別に齟齬の存在をハイライトさせた図 7 に示すように，そもそも経験だけに基づいてハザードマップを確認しない住民の存在に加えて，紙版や pdf 版のハザードマップの読図で素因の有無や凡例に基づいて危険性の内容を正確に読み取ることの困難性の一端が示されたのではないかと考えている。

また，図 7 からは，自分自身の知識・経験や紙のハザードマップを根拠とする場合と比較して，防災アプリおよび市区町村からのチラシを根拠とする場合の方が，齟齬の発生率が低いことが窺える．このため，防災アプリの普及ならびにさらなる機能強化，および市区町村からの固定資産税納入通知書へのリスク記載などの方法で，住民に正しく自然災害のリスクの有無と有りの場合の具体的内容の認識の徹底化が期待されるところである．

なお，認識の齟齬については，リスクなしとの認識が実際はリスクありと認識リスクの内容よりリスク大（より深い浸水深など）という過小認識および認識リスクの内容よりリスク小とリスクありとの認識が実際はリスクなしという過大認識に分かれるが，過小認識の場合は必要な立退き避難行動につながらない問題があるのに対して，過大認識の場合には，避難行動が不要な場合に立退き避難行動で被災する可能性があるという問題があり，適切にリスクを認識した上での在宅避難を含めた避難行動が求められることに留意する必要がある．

5．ウェブ調査による防災アプリの有効性評価

我々が開発を進めている防災アプリ「ハザードチェッカー」が提供する機能や情報については，図 8 に示すように，「役に立つ」と「やや役に立つ」を併せて有用評価と考えると，素因に関する危険性表示では 60%を超える評価が得られ，誘因に関する情報表示では，素因がない場合には関心が低くなるためか，50%前後の評価となっている．現在地や指定した地点を中心に配置して表示する各種のハザードマップは 65%を超え

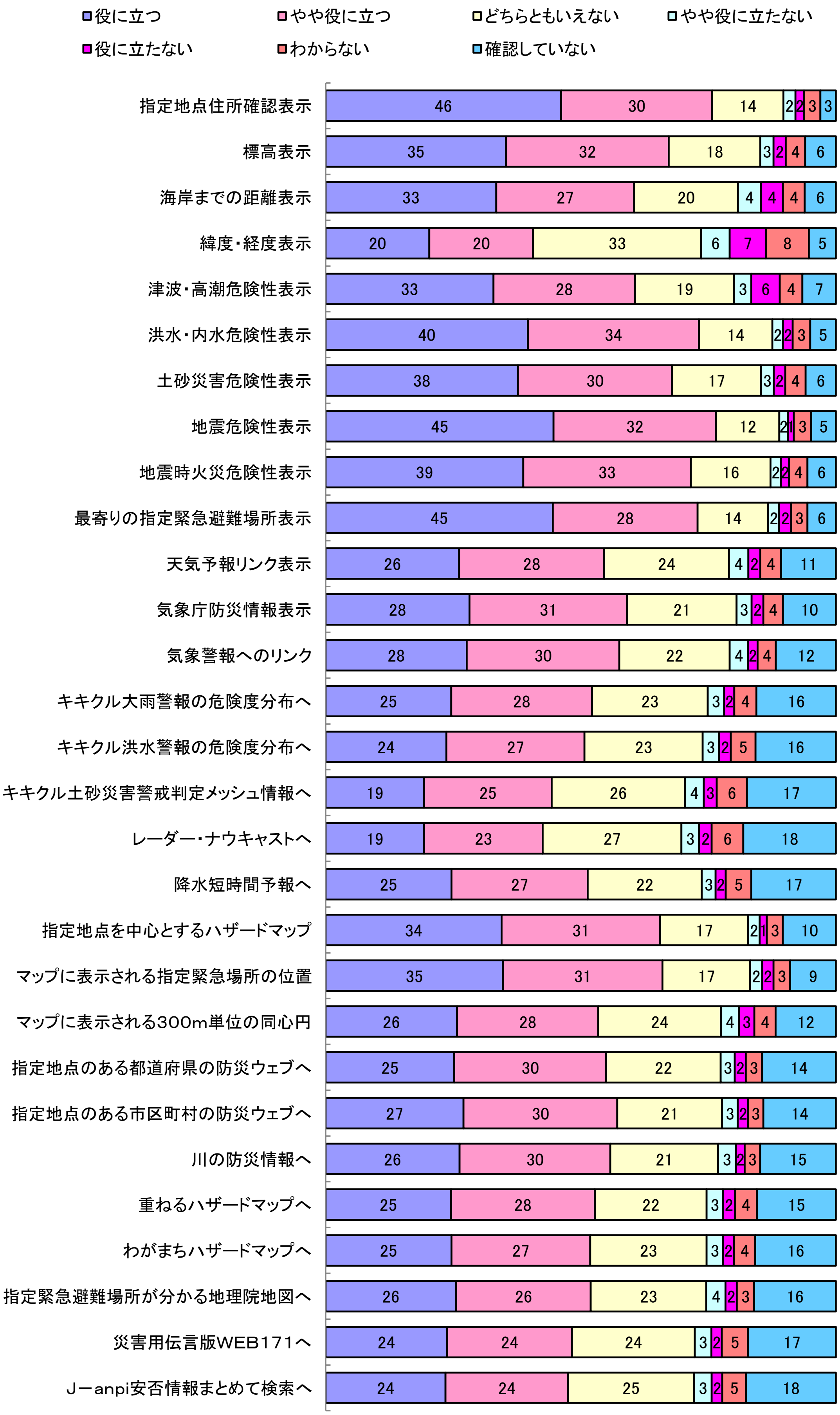


図 8　防災アプリ「ハザードチェッカー」の 29 の提供機能の利用者評価

る高い評価が得られている．ただし，PC では画面が大きくて一覧表示で多くの情報を取得できるが，画面が小さいスマートフォンでの利用では表示順が後になる機能や情報については，その機能や情報を「確認していない」との回答の比率が高くなってきており，表示する機能や情報の配置と情報間の連携関係の再検討が必要となっている．

また，「ハザードチェッカー」に対する自由回答による評価（回答者 400 人）については，有用であるという評価の一方で，文字情報が多すぎて一見して見るのをやめてしまうなどの評価もあり，機能および情報の表示方法について，さらなる検討が必要であるとの示唆が得られている．

６．おわりに

本稿では，減災に向けて，自宅の自然災害に対する脆弱性の存在を住民が正しく認識しているかどうかについて，ウェブ調査では先例がないと思われる，調査の途中で防災アプリにアクセスして認識に齟齬がないかどうかを確認してもらう，3 段階から構成される方法で調査を実施し，自然災害リスクが正しく認識されているかどうかの現状を把握することを試み，認識に齟齬がある場合の原因について考察した．

しかし，本稿での分析は，単純集計ならびに簡単なクロス集計分析に基づくものにとどまっており，齟齬の有無を被説明変数とし，個人属性および住居属性やリスクの有無の判断根拠となる媒体などを説明変数とするロジスティック回帰分析の適用による分析を行うことなどが今後の課題となっている．

また，ソフト防災に資する防災アプリの利用シーンは，住民が居住地で使用する場合と，旅行などの訪問地で使用する場合に大別される．前者の場合は，ハザードマップでハザードの有無と内容を確認し，防災・避難情報に基づくマイタイムラインを作成しておくことが推奨されているが，本研究で開発を進めている防災アプリは警戒レベルに達した際に自動的に受信されるプッシュ型ではなく，利用者自らがアクセルするプル型となっており，今後はプッシュ型のアプリの開発も検討していきたい．

謝辞 : 本発表は，科学研究費補助金（19K04884 :「ひとりひとりに届いて心配性バイアスを惹起する危機対応ナビゲーターの構築」と 20K05031 :「ソフト防災に資する防災情報の情報品質の向上と自主防災組織の活性化に関する研究」）の助成を受けた研究の成果の一部である．

参考文献

有馬昌宏（2017），ソフト防災に果たす防災アプリの可能性と課題，横幹，Vol.11，No.2，pp.145-155.

有馬昌宏・川向肇・阿部太郎（2023a），防災アプリ「ハザードチェッカー」の改良と利用者評価，日本災害情報学会第 26 回学会大会予稿集，pp.10-11.

有馬昌宏・川向肇・阿部太郎（2023b），ソフト防災を機能化させるために必要な防災情報とその効果的な提供方法に関する研究，2023 年地域安全学会梗概集，掲載予定.

田中健一郎・有馬昌宏（2016），ハザードマップの情報品質を高める防災アプリの開発，日本災害情報学会第 18 回学会予稿集，pp.222-223.

廣井悠・保科宗一朗（2020），避難情報の対象範囲に関する一考察，災害情報，No.18-2，pp.169-175.

Y.W. Lee, D.M. Strong , B.K. Kahn and R.Y. Wang(2002), AIMQ:a methodology for information quality assessment, Information & Management, 40, pp.133-146.

正規性の検定の実用例とSASでの解析方法

○小林　邦世

（イーピーエス株式会社）

How to check the Normal Distribution?

Kuniyo Kobayashi

EPS Corporation

要旨

統計の書籍に必ずといっていいほど記載のある正規性の検定. しかしながら正規性の検定については種類が多い一方で，実務で使用する機会が少なく省みられる機会が少ないのではないかと考える．そこで本発表では正規性の検定の種類についてまとめ, 解析における実例と, その実装の SAS での方法について紹介したい.

キーワード：正規性の検定, proc univariate，ヒストグラム，歪度・尖度

1，緒言

各種統計の書籍において正規性を確認するための正規性検定が紹介されてはいるが，実務で利用する機会は非常に少ない. 本論文では, 1標本データの正規性の検定をSASで実装する方法を改めてまとめるとともに，正規分布であるかどうかを確認するいくつかの方法をまとめ，実際のデータを用いて挙動を確認する．

2，正規分布であることを確認する方法

正規分布であることを確認する統計的手法として，以下の4個の手法が挙げられる．①ヒストグラム描画，②正規QQプロット，③歪度・尖度，④正規性の検定．以下からはこれらの手法を具体的に紹介していく．

① ヒストグラム描画

ヒストグラムとは，各階級に含まれるデータ数を表す「度数」を集計するための区間を表す「階級」を横軸に，その「度数」を縦軸にとったグラフである．ヒストグラムを用いることで，各階級に含まれるデータの数とデータの分布を大まかに把握することができるため，分布の形状を視覚的に捉えることが可能である[1]．すなわち，データが従う分布の形状が正規分布であるか否かを捉えるために，データ解析の最も初歩の段階で利用されるのがヒストグラムである．

しかしながら，可視化だけではデータの分布の形状については見た人間の主観的な判断に委ねることにな

るため，データの正確な記述や，対象の理解といった本質的な目標の達成は困難である場合もある[1]. また，階級の幅の設定次第でヒストグラムの見た目や印象が大きく変わること，スタージェスの式やスコットの公式など階級幅を設定する手法は提言されているものの，必ずしもコンセンサスを得られないこともヒストグラムの弱点として挙げられる.

以下にSASでのヒストグラムの描画の基本文法と，サンプルデータを用いて実際に描画したヒストグラムを示す[3]. なお，サンプルデータは1点からランダムに空いた複数の点の距離の長さを計測した「Distance」データとする.

```
SASでのヒストグラム描画
PROC UNIVARIATE DATA = [データセット名] ;
   HISTOGRAM [対象変数] /
   VSCALE = count
   VAXISLABEL = '縦軸のラベル'
   ODSTITLE = 'ヒストグラムのラベル'
   オプション
;
RUN;
```

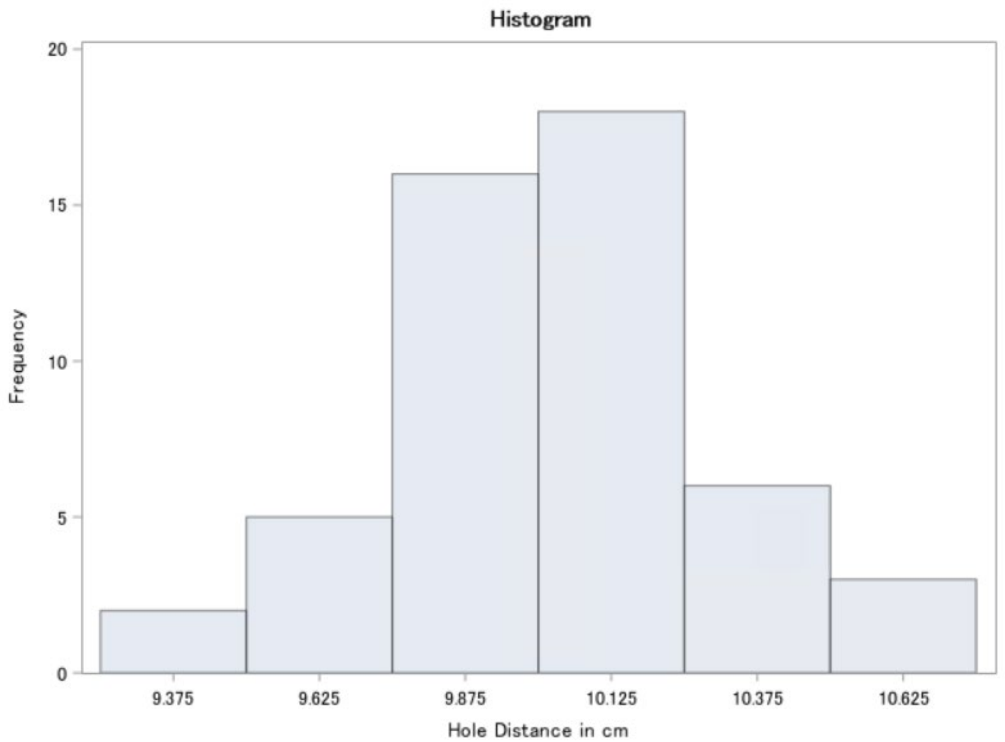


図 1　サンプルデータのヒストグラム

表 1 ヒストグラム描画の際のオプション（一部）

オプション	値	説明
vscale	count	ヒストグラムの縦軸に「度数」をとるように指定する．この指定がない場合，ヒストグラムの縦軸は各階級の「割合」がとられる．
vaxislabel	‘縦軸のラベル’	縦軸のラベルを指定する．指定がない場合，「vscale=count」の指定がある場合には’count’，ない場合には「percent」がラベルとして表示される．
odstitle	‘ヒストグラムのラベル’	ヒストグラムのラベルを指定する．指定がない場合，’Distribution :[対象変数名]’がヒストグラムのラベルとして表示される．

ヒストグラム描画の基本文法中のオプションについては表1を参照．その他のオプションについては数が膨大なため本論文では紹介を割愛する．詳しくはSAS user’s guide 9.4(2020)[3]を参照のこと．

図1に示したように，サンプルデータの分布の形状を視覚的に得ることができた．しかしながら，このデータが従う分布が正規分布であるか否かはグラフの見た人の主観に委ねられ，確かな評価とはならない．そのため，次で示す正規QQプロットや歪度・尖度，正規性の検定などを用いて定量的に評価する必要がある．

② 正規QQプロット

正規QQプロットとは観測値が正規分布に従う場合の期待値をY軸にとり、観測値そのものをX軸にとった確率プロットのこと．観測値を昇順に並べた順位から累積確率を求め，正規分布の確率密度関数の逆関数を用いて期待値を予測する．プロットが一直線上に並べば，観測値は正規分布に従っていると考えられる．

データの平均値と標準偏差をパラメータとしてもつ正規分布を仮定し，QQプロットとしてグラフにガイド線として重ねて描画することで，データの正規性を視覚的により判断しやすくする方法もある．

以下にSASでの正規QQプロットの描画方法と，サンプルデータを用いて実際に描画した正規QQプロットを示す[3]．なお，サンプルデータは①でも利用したDistanceデータとする．

```
SASでのQQプロットの作成
PROC UNIVARIATE data =[データセット名] ;
   QQPLOT [対象変数];
   QQPLOT [対象変数] / NORMAL(MU=est SIGMA=est)*
     オプション
;
(*データの平均と標準偏差をパラメータとしてもつ正規分布のQQプロットをガイド線として重ねて描画．)
RUN;
```

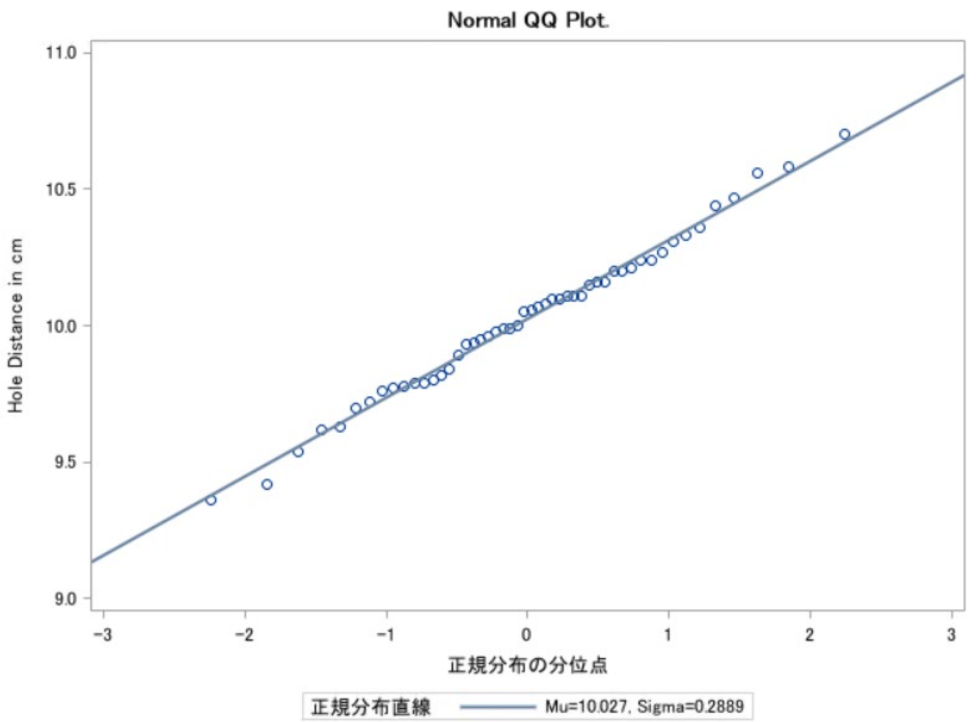


図 2　サンプルデータの正規 QQ プロット

QQプロット描画の基本文法中のオプションについては，ヒストグラム描画同様，数が膨大なため本論文では紹介を割愛する．詳しくはSAS user's guide 9.4(2020)[3]を参照のこと．

図2で示されたサンプルデータの正規QQプロットより，プロットが大まかに1直線に並んでおり，ガイド線として描画された平均10.027，標準偏差0.2889の正規分布のQQプロットとも概ね重なっていることから，サンプルデータの従う分布は概ね正規分布であると主張できる．しかしながら，正規QQプロットにおいても視覚的に判断せざるを得ず，ある程度図を見た人の主観性に判断が委ねられる部分も少なくない．

したがって，さらに機械的な正規性の評価方法を以下に示していく．

③ 歪度と尖度

歪度とは，分布が正規分布からどれだけ歪んでいるかを表す統計量で、左右対称性を示す指標のこと．ここでn個の観測値からなるデータの各観測値数を$x_1, \cdots, x_n$，データの平均値を$\bar{x}$，標準偏差をs，歪度をα_0で示すと，α_0は以下の式で与えられる．

$$\alpha_0 = \frac{n}{(n-1)*(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

正規分布であればα_0は0をとり，任意の分布のα_0が$\alpha_0 > 0$であればその分布の右の裾が長く，$\alpha_0 < 0$であればその分布の左の裾が長くなる．

また，尖度とは，分布が正規分布からどれだけ尖っているかを表す統計量で、分布の山の尖り度と裾の広がり度を示す．歪度同様，尖度をα_1で示すと，α_1は以下の式で与えられる．

$$\alpha_1 = \frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s})^4 - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$

正規分布の尖度が3であることを利用し，(任意の分布の$\alpha_0 -$ 3)が0より大きければ正規分布よりも尖っており，0より小さければ正規分布よりも丸く鈍い形をしていると表現される[1].

上記で示したように，正規分布であれば歪度は0，尖度は3であることを利用し，尖度もしくは歪度から正規性を検定することも可能であるが，SASでの実装がないことから本論文では紹介を割愛する．なおRでは尖度を用いた正規性の検定である「尖度のダゴスティーノ検定」の実施が可能である(詳細は付録2を参照).

以下にSASでの歪度・尖度の算出方法と，サンプルデータを用いて実際に算出した歪度・尖度の結果を示す[3]．なお，サンプルデータは①でも利用したDistanceデータとする．

```
SASでの歪度・尖度の算出
PROC UNIVARIATE DATA =[データセット名] ;
   VAR [対象変数];
RUN;
```

モーメント			
N	50	重み変数の合計	50
平均	10.0268	合計	501.34
標準偏差	0.28892793	分散	0.08347935
歪度	-0.0067141	尖度	0.03117379
無修正平方和	5030.9264	修正済平方和	4.090488
変動係数	2.8815567	平均の標準誤差	0.04086058

図 3 SAS における歪度・尖度の算出結果 (上から 3 行目参照)

図3で示された歪度・尖度の算出結果より，歪度が-0.007，尖度が0.003であるという結果が得られた．正規分布であれば歪度は0，尖度は3であることを利用すると，サンプルデータが従う分布は対称性においては正規分布と同等程度であるが，山は正規分布よりも丸く鈍い形をしていると解釈できる．したがって，サンプルデータの従う分布は概ね正規分布である，と主張することができる．

しかしながら尖度においてやや正規分布よりも小さい点を無視することができない．そのため，正規性をより強く主張するために，次に正規性の検定について紹介していく．

④ 正規性の検定とは

正規性の検定の代表的な手法として，コルモゴロフ・スルノミフ検定(以下KS検定)とシャピロ・ウィルク検定(以下SW検定)の2種類が挙げられる．SW検定はサンプルサイズが2000以下の場合のみ使用可能である[4]．KS検定，SW検定ともに帰無仮説に「(任意の分布は)正規分布である」ことを置くため，帰無仮説が棄却されると任意の分布は正規分布「でない」ことが示される．

KS検定は，Andrey KolmogorovとNikolai Smirnovによって提案された統計的解析手法である．KS検定では2標本の母集団の確率分布が異なるものであるかを検定するものであるため，1標本の場合には任意の分布の比較対象は正規分布のみに留まらない．1標本の場合に比較対象を正規分布に限定する形で改良されたLilliefors検定も存在するが，ここではKS検定の手順について以下に簡単に紹介する．

1, 帰無仮説を「標本Ｘが正規分布の確率密度関数から発生」とする．

2, 標本Xの累積確率分布と確率正規分布の密度関数の累積確率分布を求める．

3, 上記の2つの累積確率分布の差の絶対値の最大値であるKS統計量Dを求める．

4, 標本個数nと上記のKS統計量Dを用いて$D\sqrt{n}$の値を計算する．

5, 有意水準5%の場合，$D\sqrt{n}$の値が1.36 以上であれば帰無仮説を棄却し，「標本Xの従う分布は正規分布に一致しない」と結論付けることが可能[4][5].

KS検定では，統計量は分布の裾の部分よりも中央値付近の方に強く依存し，多くのサンプルサイズを必要とする(サンプルサイズが20以上でないと精度がやや落ちる[6]).

SW検定は，Samuel Sanford ShapiroとMartin Wilkによって提案された統計手法である．SW検定ではサンプルデータの集合の順序統計量と正規分布の順序統計量の相関(共分散)を利用した検定統計量Wを用い，検定統計量Wが小さい，すなわち，データの順序統計量と正規分布の順序統計量の相関が弱い場合に帰無仮説が棄却される．

過程については複雑なため，ここでは詳細な説明は省く．詳細についてはShapiro SS, et al. (1965)[7]を参照のこと．

SASではKS検定，SW検定のほかにAnderson-Darling検定およびCramer-von Mises検定の4種類の検定結果が一度に出力される．以下にSASでの正規性の検定の実装方法と，サンプルデータを用いて実際に検定を実

施した結果を示す[3]．なお，サンプルデータは①でも用いた「Distance」データとする．

```
SASでの正規性の検定
PROC UNIVARIATE DATA =[データセット名]　NORMAL;
  VAR [対象変数];
RUN;
```

正規性の検定				
検定	統計量		p 値	
Shapiro-Wilk	W	0.992264	Pr < W	0.9845
Kolmogorov-Smirnov	D	0.051999	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.02621	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.162071	Pr > A-Sq	>0.2500

図 4　SAS での正規性の検定の結果

図4で示された正規性の検定の結果より，KS検定，SW検定，Anderson-Darling検定，Cramer-von Mises検定の4種類の検定結果全てにおいて，帰無仮説が「保留」されている．したがって，サンプルデータの従う分布は正規分布とは異ならないと主張することができる．

⑤ 正規性の検定はどのような場面で利用されるか

臨床試験登録サイトClinicalTrials.Govにて，”Kolmogorov-Smirnov”，”Shapiro-Wilk”と検索したところ，それぞれ73試験，48試験が該当した．例えば，多囊胞性卵巣症候群患者のDiane-35とメトホルミンの併用治療におけるプエラリンの追加の治療効果を評価した臨床試験において，KS検定がデータの分布の正規性を評価されるために使用されていた[8]．また，糖尿病患者における口腔洗浄に対するショウガ洗口液，アロエベラ洗口液，生理食塩水でのマウスウォッシュの効果を比較評価した臨床試験において，データの分布の正規性の評価をSW検定を用いて実施していた[9].

3, 実際のデータを利用し，各手法で正規性を確認する

データ分析の第一歩として，データを特徴づけるために平均値や標準偏差，歪度・尖度といった代表値の算出やヒストグラム，箱ひげ図，ドットプロットなどの描画が推奨される場合が多い．その場合にはある程度作図をし，代表値を算出した上で分布の形を確認したところで解析に着手するパターンが少なくはないが，目視と代表値だけの確認で正しく分布の形を把握できているのだろうか.

以下に実際のデータについて作図，代表値，検定の順番で確認していく.

奈良県奈良市(日本の近畿地方にある当道府県の一つの県庁所在地)の2018年から2022年の7月の平均気温のデータ[10]について正規性を確認する．確認方法としては①ヒストグラム描画，②正規QQプロット描画，③歪度・尖度の確認，④正規性の検定の順番で実施する．

データの平均値は26.9，標準偏差は2.3である．

① ヒストグラム描画の結果

図5は奈良県奈良市の2018年から2022年の7月の平均気温のデータをヒストグラムで描画したものである．一見対称性のある山なりの分布であり，正規分布のように見受けられる．しかしながら，やや分布の山の左側の裾が広くなっているようにもみられ正規分布と断定して良いか判定が難しい．

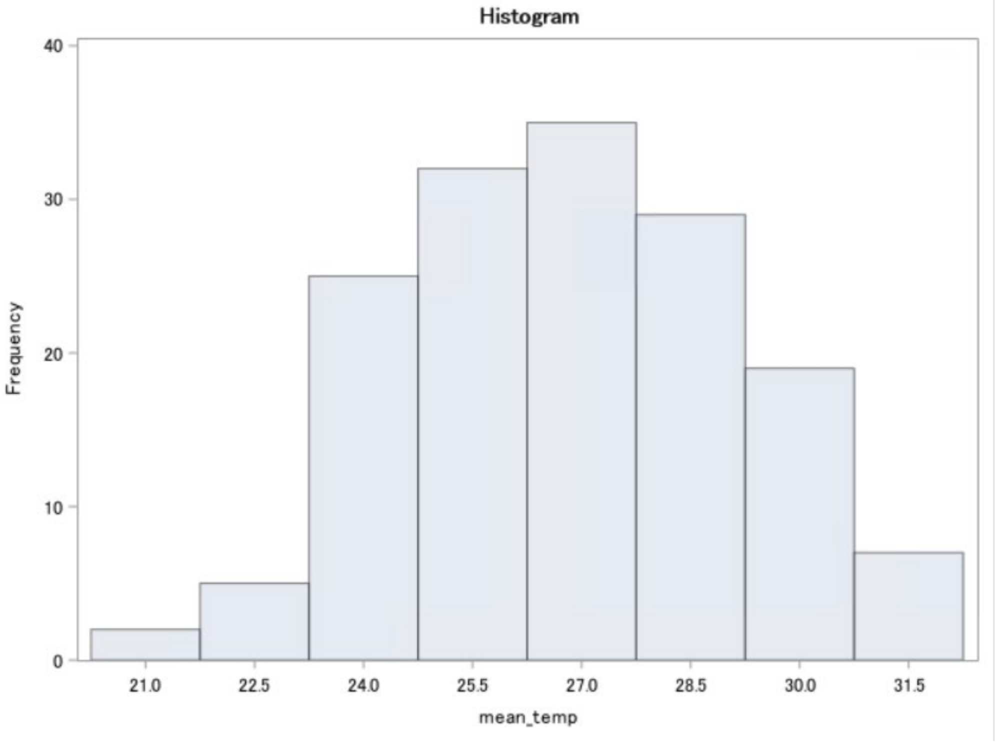


図 5 平均気温のデータのヒストグラム

② 正規QQプロット描画の結果

図6は奈良県奈良市の2018年から2022年の7月の平均気温のデータを正規QQプロットで描画したものである．データの平均値と標準偏差をパラメータにもつ正規分布の正規QQプロットのガイド線に概ね重なっているため，本データは正規分布である可能性が高い．しかしながら，平均値から離れるにつれてガイド線から外れるデータが増えていること，目視のみでの判断となり主観が入りやすいことから，正規分布であると断言して良いか判断が難しい．

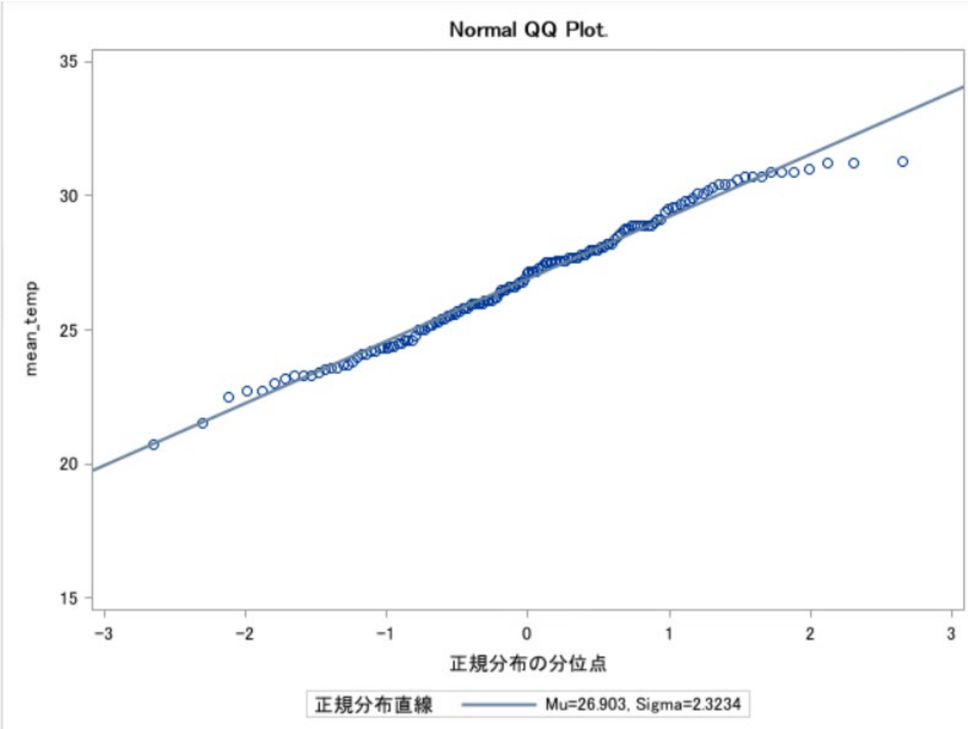


図 6　平均気温のデータの QQ プロット

③ 歪度・尖度の確認

図7は奈良県奈良市の2018年から2022年の7月の平均気温のデータの歪度・尖度の算出結果である．歪度が-0.092，尖度が-0.6233であるという結果が得られた．正規分布であれば歪度は0，尖度は3であることを利用すると，データが従う分布はやや非対称であり，山は正規分布よりも丸く鈍い形をしていると解釈できる．しかしながら分布の形を数値的に比較することは可能であっても，正規分布か否かを判断することは困難である．

モーメント			
N	154	重み変数の合計	154
平均	26.9032468	合計	4143.1
標準偏差	2.32343612	分散	5.3983554
歪度	−0.0924051	尖度	−0.6233495
無修正平方和	112288.79	修正済平方和	825.948377
変動係数	8.63626662	平均の標準誤差	0.18722782

図 7　平均気温のデータの歪度・尖度の結果（上から 3 行目）

④ 正規性の検定の結果

図8は奈良県奈良市の2018年から2022年の7月の平均気温のデータの正規性の検定結果である．各検定ともに有意水準を0.05に設定したところ，すべての検定においてp値が0.05を上回るという結果になった．したがって，奈良県奈良市の2018年から2022年の7月の平均気温データの分布は正規分布とは異ならないという結論が導き出された．

正規性の検定				
検定	統計量		p 値	
Shapiro-Wilk	W	0.984853	Pr < W	0.0905
Kolmogorov-Smirnov	D	0.055895	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.064758	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.486825	Pr > A-Sq	0.2289

図 8　平均気温のデータの各種正規性の検定結果

4, 結論

本論文では1標本データの分布が正規分布であるかどうかを確認するいくつかの手法を，SASでの実装例を伴いながら紹介した．データの分布の正規性の確認方法としてヒストグラムやQQプロットでの描画が代表的な手法に挙げられるが，正規性の検定などの他の手法も同時に使用することで，多面的な視点から分布の形を推定することが可能だと考えられる．

5, 参考文献

[1] 東京大学出版，統計学入門, 1991, 18-28

[2] 東京大学出版，自然科学の統計学, 1992, 223-225

[3] Base SAS® 9.4 Procedures Guide Statistical Procedures Sixth Edition, 2020, 289-548

[4] Kolmogorov A， "Sulla determinazione empirica di una legge di distribuzione", G. Ist. Ital. Attuari, 1933, 4: 83–91.

[5] Smirnov N, "Table for estimating the goodness of fit of empirical distributions". Annals of Mathematical Statistics. Ann. Math. Statist, 1948 19(2): 279-281

[6] Anthony J. B et.al, Informal versus formal judgment of statistical models: The case of normality assumptions, Psychonomic Bulletin & Review, 2021, 28:1164–1182

[7] Shapiro SS, Francia R. An approximate analysis of variance test for normality. Journal of the American Statistical Association. 1972; 67(337):215–216.

[8] Li W, Hu H, Zou G, Ma Z, Liu J, Li F. Therapeutic effects of puerarin on polycystic ovary syndrome: A randomized

trial in Chinese women. Medicine (Baltimore). 2021 May 28;100(21).

[9] Badooei F, Imani E, Hosseini-Teshnizi S, Banar M, Memarzade M. Comparison of the effect of ginger and aloe vera mouthwashes on xerostomia in patients with type 2 diabetes: A clinical trial, triple-blind. Med Oral Patol Oral Cir Bucal. 2021 Jul 1;26(4).

[10] 国土交通省 気象庁，過去の気象データ・ダウンロード
https://www.data.jma.go.jp/risk/obsdl/index.php

付録1：サンプルデータ「Distance」コード

```
data Distance;
   input Distance @@;
   label Distance='Hole Distance in cm';
   datalines;
 9.80 10.20 10.27  9.70  9.76
10.11 10.24 10.20 10.24  9.63
 9.99  9.78 10.10 10.21 10.00
 9.96  9.79 10.08  9.79 10.06
10.10  9.95  9.84 10.11  9.93
10.56 10.47  9.42 10.44 10.16
10.11 10.36  9.94  9.77  9.36
 9.89  9.62 10.05  9.72  9.82
 9.99 10.16 10.58 10.70  9.54
10.31 10.07 10.33  9.98 10.15
;
run;
```

付録2：尖度のダゴスティーノ検定のRでの実装方法

SASでは歪度・尖度を用いた正規性の検定，「歪度(尖度)のダゴスティアーノ検定」の実装はされていないが，Rでは尖度を利用した「尖度のダゴスティーノ検定」の実装がなされている．そこで，本付録ではRでの「尖度のダゴスティアーノ検定」の検定方法について紹介したい．

Rでの実装方法は「moments」パッケージ内の「agostino.test」関数を利用する.

```
install.packages("moments")      #momentパッケージをインストール
library(moments)                 #momentaパッケージをライブラリから読み込み
agostino.test（[変数名]）         #尖度のダゴスティーノ検定の実施
```

以下はサンプルデータ「Distance」を用いたRでの実施例と結果である.

```
install.packages("moments")
library(moments)
Distance <- c(9.80, 10.20, 10.27, 9.70,  9.76,
             10.11, 10.24, 10.20, 10.24, 9.63,
              9.99,  9.78, 10.10, 10.21, 10.00,
              9.96,  9.79, 10.08,  9.79, 10.06,
             10.10, 9.95,  9.84, 10.11,  9.93,
             10.56, 10.47, 9.42, 10.44, 10.16,
             10.11, 10.36, 9.94,  9.77,  9.36,
              9.89,  9.62, 10.05, 9.72,  9.82,
              9.99, 10.16, 10.58, 10.70, 9.54,
             10.31, 10.07, 10.33,  9.98, 10.15)
agostino.test(Distance)
```

```
        D'Agostino skewness test

data:  Distance
skew = -0.006511, z = -0.020957, p-value = 0.9833
alternative hypothesis: data have a skewness
```

p値が有意水準0.05を大きく上回るため帰無仮説が保留され，サンプルデータ「Distance」の分布は正規分布と異ならないと主張できる．これにより，SASで実施した4種類の「正規性検定」と結果は同様となっていることが確認できた.

世界価値観調査（WVS）に見る
日本人の最近40年間の価値観推移

武藤　猛

（MarkeTech Consulting）

Transition of Japanese Values during recent 40 years based on WVS

Takeshi Muto
MarkeTech Consulting

要旨

世界価値観調査（World Values Survey=WVS）は 1981 年から行われている、日本を含む延べ 100 か国以上が参加する価値観に関する大規模社会調査である。既に 7 回実施され、日本は初回から参加している。本論文では因子分析により日本人の主要な価値観を抽出しその推移について考察した。分析の方法には、(1)WVS の HP にある Online Analysis を利用して Excel でダウンロード（D/L）した集計表レベルの分析、(2)全時系列データを一括 D/L した個票レベルの分析、の 2 種類がある。まず集計表レベルの分析として、「Inglehart-Welzel Map」の再現を試みた。この Map は、世界各国を 2 次元の Map に位置付けたものである。WVS の HP に記載されている 10 個の変数を用いて因子分析を行い、ほぼ完全に上記 Map が再現できることを確認した。つぎに日本に限定した個票レベルの分析を行った。WVS の質問数は最大 265 と多いので、まず各回に含まれかつ価値観に直接関連した質問 44 から出発し、多段階因子分析による変数の刈り込みを行って最終的に 22 変数を得た。因子分析の結果、1.国への誇り、2.権力への従順さ、3.規範意識の強さ、4.社会への信頼、5.生活への満足感、という 5 因子を得た。これらの因子の 40 年間の推移について考察した。

キーワード：WVS、価値観、社会調査、因子分析

１．　世界価値観調査（WVS）と価値観の基本的な考え方

（１）世界価値観調査（WVS）の概要

世界価値観調査（World Values Survey=WVS）は、世界規模で実施されている社会調査であり[(1)]、ミシガン大学社会調査研究所のロナルド・イングルハート(1934 -2021)らが創始した。既に創設されていた European Values Study（EVS）と連携した国際プロジェクトである。主催者は World Values Survey Association（本部：スウェーデン・ストックホルム）である。WVS は 1981 年以来約 5 年おきに、既に 7 回実施されている。WVS には延べ 100 以上の国や地域が参加しており、日本は初回から参加している[(2)]。WVS の社会調査としての概

要を図表 1 に示す。なお、WVS では各回の調査を Wave と呼び、各回を W1～W7 と略称する。また日本の場合は、W1(81)のように、調査実施年の下二桁を括弧内に追記する。

図表1. 世界価値観調査(WVS)の概要

調査(Wave)	WVS全体			日本	
	実施年	参加国／地域数	個票件数	実施年	個票件数
W1	1981-1984	11	14,840	1981	1,204
W2	1990-1994	21	29,174	1990	1,011
W3	1995-1998	55	77,818	1995	1,054
W4	1999-2004	41	60,045	2000	1,362
W5	2005-2009	58	85,149	2005	1,096
W6	2010-2014	60	89,565	2010	2,443
W7	2017-2022	64	94,278	2019	1,353
		(延べ) 108	(計)450,869		(計) 9,523

WVS のデータ利用については、WVS のウェブサイトで全データが公開されている。データは次の二つの方法で利用可能である：

① 上記ウェブサイトの「Online Analysis」ページから Wave 別・国別・質問別の集計表を作成する

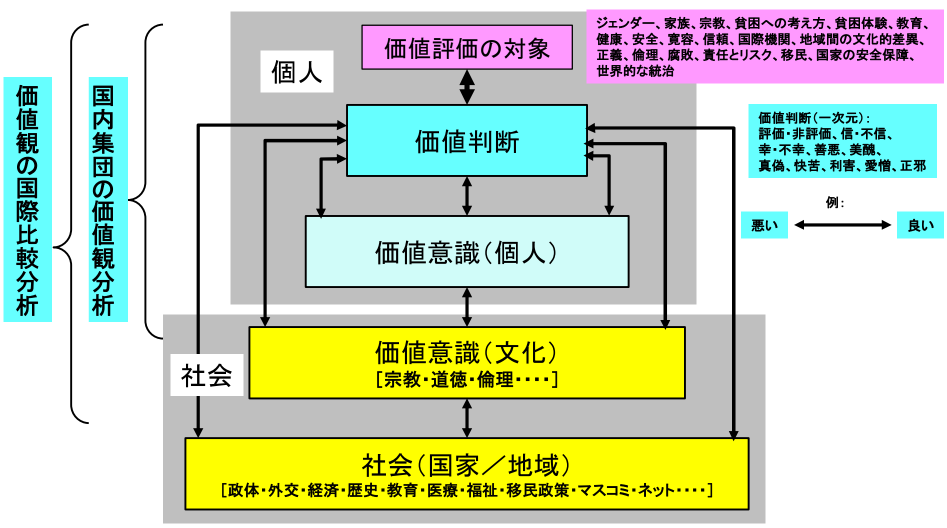
② データを一括 D/L して分析プログラムにより分析する

（２）価値観の基本的な考え方

WVS の調査内容は、大部分が価値観に関するものである[(3),(4),(5)]。価値観とは、「何に価値を認めるかという考え方。善悪・好悪などの価値を判断するとき、その根幹をなす物事の見方」（広辞苑）である。価値観は、個人や集団の行動に大きな影響を及ぼす。つまり価値観が近いかどうかが、個人の集団化やその離散を駆動するのである。最悪の場合には、価値観の違いが集団間の紛争や戦争をもたらす場合もある。WVS が長期間、継続的に実施されている背景には、社会における価値観の重要性があると考えられる。

価値観分析の概念図を図表 2 に示す[(6)]。WVS の価値観調査は、図表 2 の「価値評価の対象」と「価値判断」とを調査票として展開したものである。「価値評価の対象」は、ジェンダー、家族、宗教、貧困への考え方、等実に多様であり、対応する「価値判断」も、評価・非評価、信・不信など様々である。ここで、「価値判断」が必ず一次元的（例：良い⇔悪い）であることが重要である。このことにより、価値観調査で多段階式により回答を求める調査が可能になる。

図表2. 価値観分析の概念図

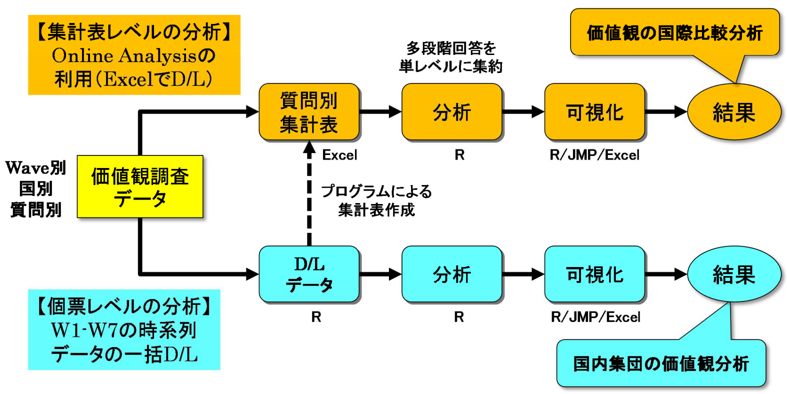


２．分析のためのデータ準備

（１）データ準備の二つの流れ

図表 3 にデータ準備の流れを示す。データ準備には二つの方法がある。まず集計表レベルの分析であり、これは WVS の HP にある Online Analysis を利用する。HP 上で Wave と調査項目、国名を選択すると、該当の集計表が Excel で D/L できる。一方、個票レベルの分析では、W1～W7 の時系列データの一括 D/L を利用する。

図表3. データ準備の流れ



（２）集計表レベルの分析のためのデータ準備

Online Analysis のページから D/L した集計表には、因子分析等で利用するために工夫が必要である。集計表レベルの分析では国間の比較が主な目的である。このため、例えば、ある質問 W7/Japan/Q1「Important in life: Family」を例に取ると、Online Analysis のページで、上記表示データを Time Series として D/L する。次に国別データを単一指標化するため、D/L したデータのうち、上位 2 段階（4 段階質問のうち、「重要」および「やや重要」）を合計してその比率を分析に使用する。他の質問の加工も同様である。

（３）個票レベルの分析のためのデータ準備

WVS データは Wave 単位でも D/L 可能であるが、Time Series データとして全 Wave を一括 D/L するのが便利である。ここで、WVS は質問項目のタイプが多様であることに注意が必要である。図表 4 は質問タイプ別に質問数を示したものである。本分析で使用するのは、価値観に関する多段階（2 段階から 10 段階まで）質問のみである。これらの質問はすべて一次元軸上で価値観を定量的に測定している。

欠損値については、狭義の欠損値（ランダムに発生し、比較的少数）と欠測値（Wave により欠測している質問）の 2 種類がある。両者とも欠損値は、すべて列平均値（変数別平均値）に置き換えることとした。また分析で使用する質問については、W1～W7 すべてで計測されかつ価値観に関連した質問 44 個に限定した。

図表4．質問タイプ別質問数

質問タイプ		質問数（および小計）		備考
段階式	2段階	32	186	価値観に関する質問（因子分析で使用）
	3段階	8		
	4段階	76		
	5段階	17		
	8段階	2		
	10段階	51		
段階式	2段階	8	8	「〇〇は信用できるか」等
選択式	2択～9択	66	66	「教会員ですか」等
選択式	属性	5	5	年齢、性別等
	（計）	265	265	

例：5段階

0 1 2 3 4

３． Inglehart-Welzel Cultural Map の再現

（１） Inglehart-Welzel Cultural Map とは

Inglehart-Welzel Cultural Map とは、WVS の成果を可視化する図として有名である[1]。この Map は、WVS のデータを国別に集計し、因子分析することで得られた次の 2 因子：

① 生存的価値 vs 自己表現的価値（Survival vs. Self-Expression Values）（生活軸＝横軸）

② 伝統的価値 vs 世俗的価値（Traditional vs. Secular Values）（宗教軸＝縦軸）

を用いて、WVS 参加国（地域）すべてを、2 次元 Map に位置づけるものである。Inglehart-Welzel Cultural Map は、W1 以来毎回作成・公表されている。上記二軸が毎回安定して WVS の可視化に役立っていることから、因子や因子名選定が適切であることが分かる。Map における地域の分類は次の通り 8 つである：African-Islamic、Catholic Europe、Confucian、English-Speaking、Latin America、Orthodox Europe、Protestant Europe、West & South Asia。日本は Confucian（儒教圏）に属している。

（２） 因子分析による Map の再現結果

Map の再現のための分析手順は次の通りである。WVS の HP に記載されている 10 個の変数を用いた。Online Analysis により、全参加国に関する W1～W7 の上記変数に関する集計表を作成した。この集計表から、多段階の集計結果を単一指標化するために、因子との相関係数がプラスになるように、上位側または下位側の回答比率を集計した。最後に、国別に単一指標化されたデータを用いて因子分析を行った。

Inglehart-Welzel Cultural Map の再現のための因子分析結果（ただし W7 の結果）は図表 5 の通りである。図表 5 に示すように変数と因子名との対応は納得できるものである。図表 6 に Inglehart-Welzel Cultural Map の再現結果を示す。オリジナルの Map（ここでは省略。WVS の HP 参照）と対比すると、日本などいくつかの国の布置をチェックすることにより、十分な精度で Map が再現できていることを確認できる。結論として、WVS データの可視化のため、因子分析は有効で再現可能な手段といえる。

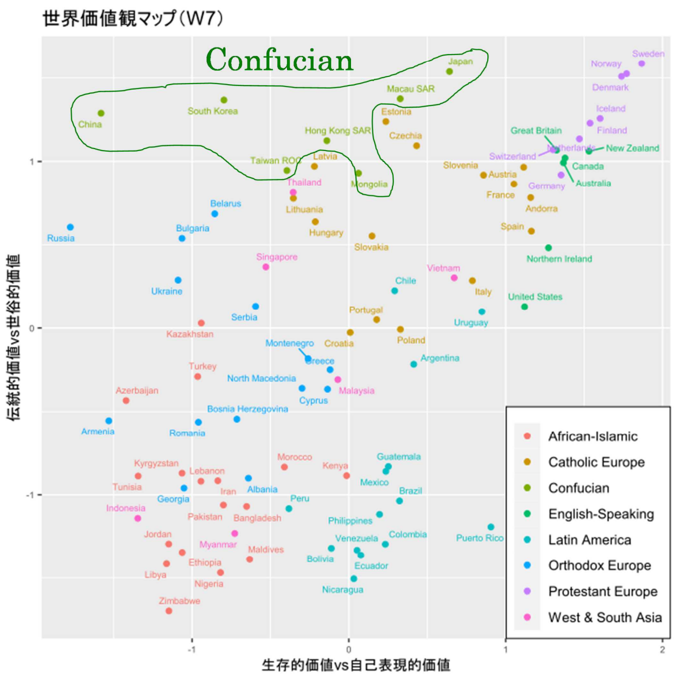
図表 6 における日本の位置付けについて付け加える。日本は Confucian（儒教圏）の中の一国として、世俗的価値を重んじ（一神教的視点からは「無神論」であるが「無宗教」ではない）、比較的豊かな国として自己表現的価値（自由）を満喫している。次節ではこのような日本人の価値観という、日本人の内実について考察する。

図表5. Inglehart-Welzel Cultural Map の再現のための因子分析結果

No.	WVS	W7	変数名	質問形式	単一指標化（代表比率）	変数の意味（－方向）	変数の意味（＋方向）	因子名	因子負荷量	
									MR1	MR2
1	F063	Q164	神への信仰	10段階	上位3段階合計(信仰)	多神教・無宗教	一神教への信仰	MR1.伝統的価値vs 世俗的価値 (Traditional vs. Secular Values)	0.92	0.07
2	A042	Q17	子供の従順さ	3段階	上位1段階(親に従順)	親への反抗	親に従順		0.76	−0.17
3	F120	Q184	権力指標	3段階	上位1段階(権力に従順)	権力への反抗	権力に従順		0.72	−0.25
4	G006	Q254	人を信頼できるか	2段階	否定(信頼できない)	人への信頼	人を信頼できない		0.68	0.25
5	Y011A	I_AUTHORITY	正当か_中絶	10段階	否定3段階合計(中絶反対)	中絶に賛成	中絶に反対		0.67	0.36
6	Y002	Y002	国への誇り	5段階	上位2段階合計(国は誇り)	グローバリズム	ナショナリズム		0.23	0.08
7	A008	Q46	物質主義指標	3段階	上位1段階(物質主義)	自由を重視	物質を重視	MR2.生存的価値vs 自己表現的価値 (Survival vs. Self−Expression Values)	−0.09	0.81
8	F118	Q182	正当か_同性愛	10段階	否定3段階合計(同性愛反対)	同性愛賛成	同性愛反対		0.31	0.71
9	E025	Q209	政治的行動_請願	3段階	否定(参加しない)	誓願に参加	誓願に参加しない		−0.01	0.71
10	A165	Q57	幸福感	4段階	下位2段階(幸福でない)	幸福である	幸福でない		0.08	0.36

［注1］W7に対する結果
［注2］WVSは一括ダウンロードデータの、W7はWave7単独の変数記号
［注3］MR1の符号については、Inglehart-Welzel Cultural Map の布置と合わせるために、正負を反転させた

図表6. Inglehart-Welzel Cultural Map の再現



4．　日本人の 40 年間の価値観推移分析

（1）分析方針

日本人の 40 年間の価値観推移の分析方針は次の通りである：

① WVS の W1～W7 データから抽出した日本のデータを用いる（方法は前述の通り）

② 因子分析により、日本の主要な価値観を抽出する。なお、予備分析により、因子数は 5 とするのが適切であることが判明している。また、Inglehart-Welzel Cultural Map の因子分析に倣って、因子別変数の数は 5 を目途とする。従って変数の合計値の目標は 5×5=25 となる。

③ 5 つの因子に適切な名称を付与する。元のデータに、5 つの因子の負荷量を付与したデータを作成し、各種分析に利用する。

（２）因子分析

前述の方針に従って、因子分析を行った。WVS の質問総数は 186 にも及び、また質問により欠測している Wave も多い。そこで、欠測 Wave がなく、かつ価値観に関する質問 44 に限定して因子分析を行うこととした。欠損値については、狭義の欠損値（ランダムに発生）よりも欠測値の方がはるかに多い。本論文では、欠損値は、すべて列平均値（変数別平均値）に置き換えることとした。まとめると、欠測 Wave 数が少なく、かつ価値観に関連した質問を用いる。結果として、分析結果に及ぼす欠損値の影響は小さいと考えられる。

因子分析については、最初の 44 変数から、目標の 25 変数に絞り込む。このため、多段階で因子分析を行う。予備解析で因子数は 5 とするのが適切であることが分かっている（スクリープロットによる）。ステップ 1 では 44 変数･5 因子で因子分析を行い、各因子に属する変数のうち因子負荷量の小さいものを「枝刈り」し、28 変数に削減した。ステップ 2 では 28 変数・5 因子で因子分析を行った。同様に、1 因子当り 5 変数を目標に「枝刈り」を行ったところ、目標の 25 変数に近い 22 変数が得られたので、この 22 変数を最終結果とした（W1～W7 のデータを使用）。22 変数・5 因子の因子分析結果を図表 7 に示す。

図表 7 に示すように、抽出された因子は次の 5 つである：「1.国への誇り」、「2.権力への従順さ」、「3.規範意識の強さ」、「4.社会への信頼」、「5.生活への満足感」。これらの命名にあたっては、図表 7 の「因子の意味に関与」欄の○印の変数名を参考にした。もちろん×印の変数も交絡等で各因子に関連していると考えられる。また各変数の意味に関して、因子をプラス方向に意味づける場合と、マイナス方向に意味づける場合とを記入してある。

図表7．日本人の40年間の価値観推移分析のための因子分析結果

WVS	W7	変数No	変数名	因子の意味に関与	変数の意味（ー方向）	変数の意味（＋方向）	因子名	因子記号	因子負荷量				
									MR1	MR2	MR3	MR4	MR5
G006	Q254	1	国への誇り	○	誇りが強い	誇りが弱い	1.国への誇り	MR1	0.97	0.00	0.00	0.01	−0.01
Y011B	I_NATIONALISM	2	ナショナリズム指標	○	ナショナリズムが強い	ナショナリズムが弱い			0.96	0.00	0.00	−0.02	0.01
F063	Q164	3	神への信仰	×	多神教･無宗教	神の信仰が篤い			−0.16	−0.06	−0.01	−0.07	0.05
E018	Q45	4	将来_権力への尊敬	○	権力へ従順	従順は良くない	2.権力への従順さ	MR2	0.00	1.00	0.00	0.00	0.00
Y011A	I_AUTHORITY	5	権力指標	○	権力へ従順	従順は良くない			0.00	1.00	0.00	0.00	0.00
E015	Q43	6	将来_仕事の重要性は減少	×	重要性低下は良い	重要性低下は良くない			−0.08	0.23	−0.08	−0.06	0.02
A035	Q12	7	子供の忍耐力	×	忍耐力は重要でない	忍耐力は重要			0.06	0.06	0.00	0.00	0.06
F116	Q180	8	正当か_税金のごまかし	○	規範意識が強い	規範意識が弱い	3.規範意識の強さ	MR3	−0.02	−0.01	0.78	0.00	0.00
F115	Q178	9	正当か_公共交通の料金を払わない	○	規範意識が強い	規範意識が弱い			−0.01	0.00	0.77	0.01	0.00
F117	Q181	10	正当か_賄賂を受ける	○	規範意識が強い	規範意識が弱い			0.00	0.00	0.70	−0.03	−0.01
F114A	Q177	11	正当か_権利のない便益の請求	○	規範意識が強い	規範意識が弱い			0.03	−0.02	0.42	0.02	0.01
F123	Q187	12	正当か_自殺	×	規範意識が強い	規範意識が弱い			0.14	0.06	0.30	0.04	−0.01
F120	Q184	13	正当か_中絶	×	規範意識が強い	規範意識が弱い			0.13	0.11	0.24	0.00	0.06
F118	Q182	14	正当か_同性愛	×	規範意識が強い	規範意識が弱い			0.16	0.14	0.22	0.05	0.11
E069_08	Q74	15	信用_公共サービス	○	信頼できる	信頼できない	4.社会への信頼	MR4	−0.02	−0.01	−0.01	0.85	−0.01
E069_07	Q73	16	信用_議会	○	信頼できる	信頼できない			0.00	0.01	−0.01	0.82	0.01
E069_06	Q69	17	信用_警察	○	信頼できる	信頼できない			0.07	0.02	0.04	0.53	−0.02
E069_04	Q66	18	信用_新聞	○	信頼できる	信頼できない			0.00	0.02	0.00	0.33	0.00
A170	Q49	19	生活満足度	○	不満	満足	5.生活への満足感	MR5	−0.01	0.01	−0.01	0.00	0.85
C006	Q50	20	家計満足度	○	不満	満足			−0.01	−0.03	−0.01	−0.03	0.67
A173	Q48	21	選択の自由	×	不満	満足			0.05	0.02	0.05	0.03	0.46
A009	Q47	22	主観的健康感	×	満足	不満			−0.02	0.01	−0.01	−0.03	−0.37

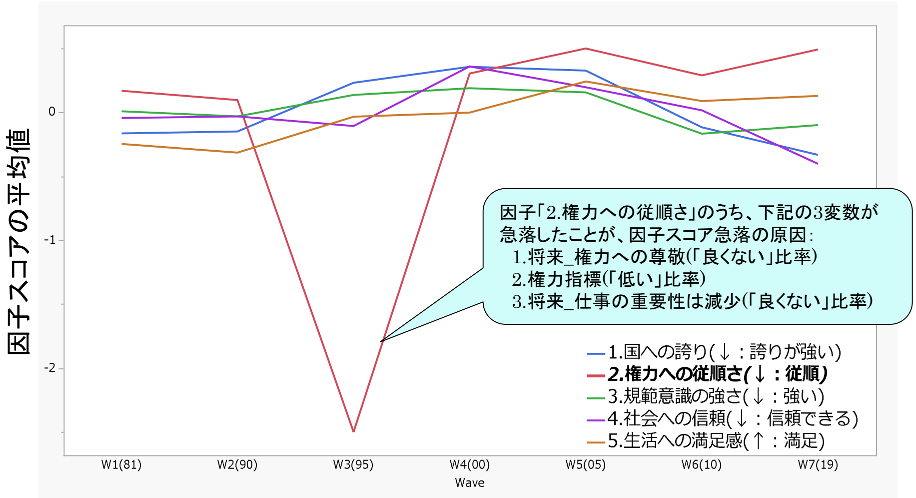
［注］WVSは一括ダウンロードデータの、W7はWave7単独の変数記号

（３）価値観推移

図表 8 は、前記の 5 つの因子で代表される価値観の約 40 年間の推移を示す。縦軸は因子スコアの平均値であり、前述のプラス方向とマイナス方向の平均値なのでその意味を議論することは困難であるが、因子 2 を除いて、穏やかに推移している。因子「2.権力への従順さ」だけが W3(95)において急落している。その原因を調べると次の 3 変数が急落したことが、因子スコア急落の原因である：「1.将来_権力への尊敬(「良くない」比率)」、「2.権力指標(「低い」比率)」、「3.将来_仕事の重要性は減少(「良くない」比率)」。これらの変数

が急落したのは、バブル崩壊時期とされる 1991 年～1993 年頃の景気後退期における政治権力や社会への不信を反映した一時的な現象と考えられる。

図表8．5つの因子で代表される価値観の推移



全ての因子には、正負の価値観がある。図表 8 はプラス方向とマイナス方向の平均値なので、次にこれらをプラス方向とマイナス方向に分離することを考える。価値評価の正負方向の推移算出方法は次の通りである：

① 「因子分析結果の詳細と因子の命名」の表で、「各因子の意味に関与」した変数のみを使用する

② 上記各変数の Wave 別集計表を作成する

③ この集計表を、価値評価に関して「－方向」と「＋方向」に二分する（各段階別回答を二分）

④ 価値評価に関する段階数が奇数の場合は、真ん中の件数を「－方向」と「＋方向」に按分する

⑤ この結果を用いて、価値評価「－方向」、および価値評価「－方向」に集約する

このような手順でまとめた結果を、図表 9 に価値評価の正負方向の推移としてグラフ化して示す。

・因子「1.国への誇り」については、「誇りに思う」が 60%～80%、「誇りに思わない」が 20%～40%で推移しており、W5(05)以降は、「誇りに思う」が増えつつある。

・因子「2.権力への従順さ」については、一時的現象である W3(95)を除くと、「良い」が 10%、「悪い」が 90%で推移しており、W3(95)を除くと変化は少ない。

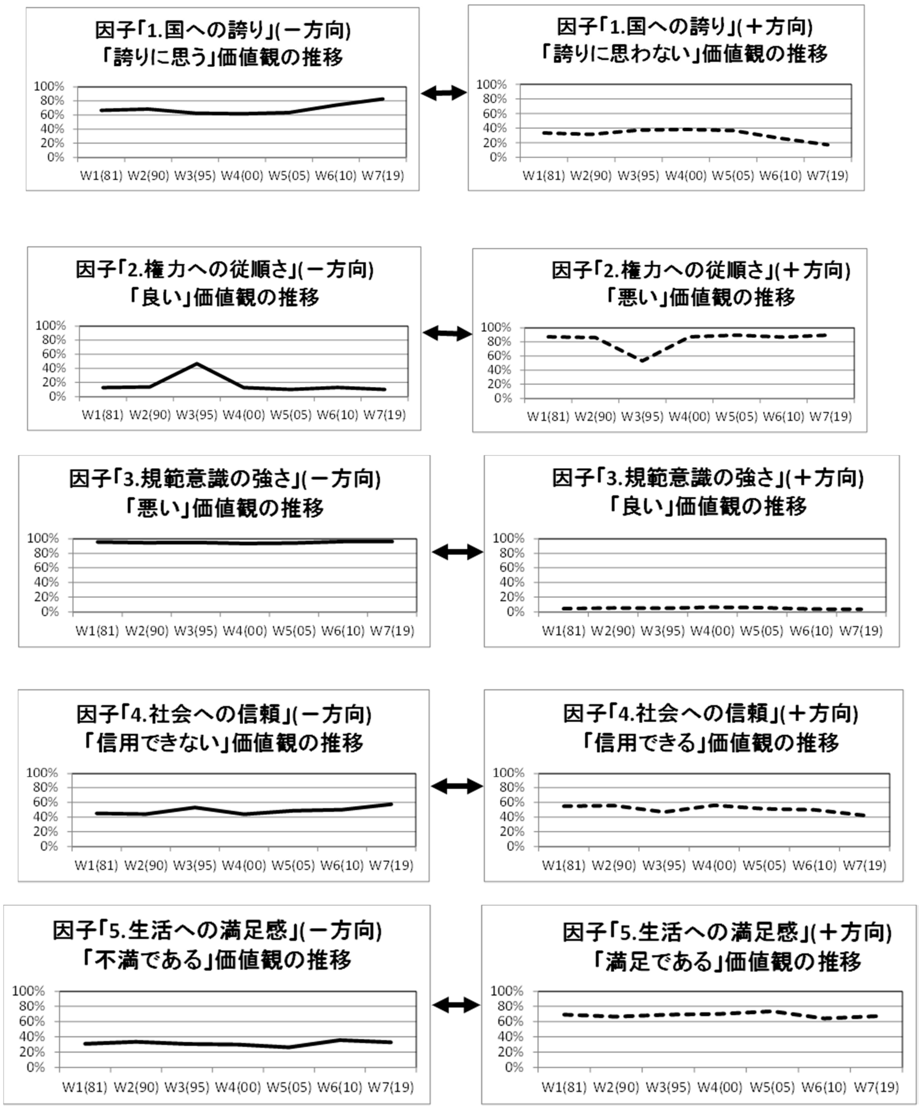
・因子「3.規範意識の強さ」については、「悪い」が 95%以上、「良い」が 5%以下で、変化は少ない。

・因子「4.社会への信頼」については、「信用出来ない」が 40%～60%、「信用できる」が 60%～40%で推移している。W4(00)以降は「信用出来ない」が漸増傾向にある。

・因子「5.生活への満足感」については、「不満である」が 20%～40%、60%～80%で推移していて、ほぼ一定範囲にある。

本分析ではセグメント別分析も行った。AGE(年齢)、SEX(性別)、EDU(教育)、INCOME(所得)、CLASS(社会階層)の 5 つのセグメントについて、価値観の推移を分析した。結果を図表 10 にまとめとして示す。ここでは AGE セグメントに関する観察結果を述べる。「1.国への誇り」は、若年層ほど誇りが弱く、高齢層ほど誇りが強い傾向がある。「2.権力への従順さ」については、年齢差は小さい。「3.規範意識の強さ」について高齢層で強く、若年層で小さい傾向がある。「4.社会への信頼」は高齢層で強く、若年層で弱い傾向がある。「5.生活への満足感」は、若年層の満足感が一貫して高い。

図表9. 価値評価の正負方向の推移



図表10. セグメント別価値観推移まとめ

セグメント｜因子名称	AGE(年齢) セグメントの動きの特徴	SEX(性別) セグメントの動きの特徴	EDU(教育) セグメントの動きの特徴	INCOME(所得) セグメントの動きの特徴	CLASS(社会階層) セグメントの動きの特徴	(まとめ)
1.国への誇り	若年層ほど誇りが弱く、高齢層ほど誇りが強い傾向がある。	ほとんどの時期で男性の方が女性よりも誇りが強いが、W6(10)以降は逆転した。	誇りの強さは、教育低セグメント＞中セグメント＞高セグメントである。	誇りの強さは、所得低セグメント＞中・高セグメントである。	社会階層が低いほど誇りが強い因傾向がある。	**「誇り」はW4(00)にかけて一時弱くなり、W4(00)以降は再び強くなった。**
2.権力への従順さ	各セグメントは、W3(95)で一旦落ち込む以外、ほぼ一定である。年齢差は小さい。	各セグメントは、W3(95)で一旦落ち込む以外、ほぼ一定である。男女差は小さい。	従順さは、教育低セグメント＞中セグメント＞高セグメントである。時期による差は小さい。	各セグメントは、W3(95)で一旦落ち込む以外、ほぼ一定である。所得による差は小さい。	各セグメントは、W3(95)で最小となり、以降はほぼフラットである。社会階層低のセグメントがより従順という傾向がある。	**W3(95)で「従順さ」が一時的に強くなり、それ以降は徐々に弱くなった。**
3.規範意識の強さ	規範意識は高齢層で強く、若年層で小さい傾向があり、ほぼ年齢順である。	すべての時期で規範意識は女性の方が男性よりも強い。	規範意識の強さは、教育低・中セグメントが強く、高セグメントが弱い傾向がある。	規範意識の強さは、所得低・中セグメントが強く、高セグメントが弱い。	W5(05)まではほぼ一定であるが、それ以降社会階層高のみが規範意識が弱まっている。	**W4(00)以前は「規範意識」はあまり変化がないが、W4(00)以降は強まっている。**
4.社会への信頼	社会への信頼は高齢層で強く、若年層で弱い傾向があり、ほぼ年齢順である。	W4(00)にピーク(信頼できない)となり、以降は信頼が強まっている。男女差は小さい。	社会への信頼は、教育低セグメントで強く、低セグメントで弱い。	一部乱高下はあるが、W4(00)以降に信頼が高まっている。セグメント間の差は小さい。	社会階層低セグメントほど社会への信頼が低い傾向がある。	**「信頼」はW4(00)以前はあまり変化がないが、W4(00)以降は強まっている。**
5.生活への満足感	若年層の満足感が一貫して高い。	満足感は一貫して女性の方が男性より高い。男女ともW5(05)以降、減少している。	満足感の高さは、教育高＞中＞低は一貫して変わらない。	満足感の高さは、所得高＞中＞低の順で一貫している。	社会階層低から高い方へと満足感が高まっている。	**W1(81)以来、多少の変動はあるが、生活への満足感が高まっている。**

５． まとめ

本論文では世界価値観調査（WVS）のデータを用いて、日本人の主要な価値観を抽出し最近 40 年間における推移を考察した。まず、WVS のシンボルとも言える「Inglehart-Welzel Map」の再現を試み、因子分析により十分な精度で再現できることを確認した。次に、WVS の W1～W7 データから抽出した日本のデータを用いて、因子分析により日本の主要な価値観を抽出し、その推移について考察した。

日本人の価値観の推移を考察するにあたり、最近 40 年間における日本の社会経済情勢の推移について確認する必要がある。

・1991 年以来、年平均経済成長率 0.9%という低成長が続いている
・出生数低下に伴う人口減少が止まらない
・高齢化の進展
・製造業の労働生産性が 1995 年の世界 1 位から、2016 年には 15 位へ低下した
・労働人口の 70%超が第 3 次産業に移行した
・2000 年以降、開・廃業率が他国の 3 分の 1 しかない
・公的教育支出が GDP 比率 2.9%で、OECD 諸国中最低（2018 年）である
・主要先進国中で、男女間賃金格差が 75.5 であり、最大レベル（2019 年）である
・経済格差の進展：ジニ係数は 0.314（1980 年）から、0.372（2017 年）に増大した

要約すれば、日本では国力の衰退が 40 年間続き、他の国々との差が開いたと言えよう。その原因については、経済政策や教育などに原因を求める様々な説がある[(7),(8),(9)]。

一方、この 40 年間の日本人の価値観には劇的な変化は見られない（一時的な変化を除く）。日本人の大多数の価値観を要約すると（図表 9 による）、国を誇りに思い、権力に従順であることを良しとせず、規範意識の強さを良しとせず、社会への信頼・不信は相半ばし、大部分が生活に満足している。日本人の生活満足度が高いという WVS の結果は、日本人が最近 40 年間の社会経済の停滞を（たとえ不本意であるにせよ）受け入れているようにも見える。なお、日本人の幸福度については、G7 諸国中で最低というデータ（OECD Better Life Index 2017）[(10)]がある。その理由として、低成長が続き生活の豊かさが感じられない、社会保障など社会的支援の低迷、官僚や政治の世界における腐敗、などが指摘されている。

日本人の特徴として巷間でよく言われる「同調圧力に弱い」や「権力に従順である」などの特徴は、上記因子分析の結果を参照すると、（WVS の調査結果に基づく限り）必ずしも大多数の日本人の特徴とまでは言えない。バブル崩壊後の日本は、リスク回避、コスト削減など後ろ向き施策に頼った「衰退途上国」の道を歩み続けた[(11)]。停滞を打ち破る動きはまだ明確ではないが、外国投資の積極的な受け入れ、同質社会を打ち破る「異分子」の活用などによる社会の再活性化などがキーとなりそうである。最近の選挙における多数の女性の首長・議員の出現が社会の変革につながるかもしれない。

参考資料

(1) WVS Database: https://www.worldvaluessurvey.org/

(2) 電通総研：世界価値観調査、https://institute.dentsu.com/keywords/wvs/

(3) ロナルド・イングルハート 、山﨑聖子訳：宗教の凋落? 100 か国・40 年間の世界価値観調査から、勁草書房 、2021

(4) ロナルド・イングルハート、山崎聖子訳:文化的進化論　人びとの価値観と行動が世界をつくりかえる、勁草書房 、2019

(5)電通総研，池田謙一 編：日本人の考え方 世界の人の考え方 II: 第 7 回世界価値観調査から見えるもの、勁草書房 、2022

(6) 見田宗介：定本 見田宗介著作集　８－社会学の主題と方法、岩波書店、2012

(7) 橘木俊詔：日本の構造、講談社現代新書、2021

(8) 森永卓郎：ザイム真理教、三五館シンシャ、2023

(9) のぶ：学校というブラック企業　元公立中学教師の本音、創元社、2023

(10) OECD Better Life Index: https://www.oecdbetterlifeindex.org

(11) NHK スペシャル：ジャパン・リバイバル“安い３０年”脱却への道、2023 年 4 月 2 日放送

大規模言語モデル狂想曲

〇小野　潔[1]

([1]コムチュア株式会社　デジタルイノベーション本部)

Large-Scale Language Model Rhapsody

Kiyoshi Ono
Digital Innovation Dept., COMTURE Corporation

要旨

2022年11月にOpenAIから発表されたChatGPTは、大規模言語モデル（LLM）の新しい時代を切り開きました。その後、Microsoft、Google 、Amazon、Metaなど、各社が次々と大規模言語モデルを発表され、その性能は急速に向上している。本論では、大規模言語モデルの発展史やアルゴリズムを通じて、その特徴を分析し、同時に日本語LLMの現状を報告する。またLLMの基礎となるBERTはSASでも稼働するので報告する。

キーワード：大規模言語モデル（LLM）、Chat-GPT、Attention、Transformer、BERT

1　はじめに

1.1　生成AI、基盤モデル、大規模言語モデル(Large Language Model, LLM)とは

近年、機械学習や深層学習の技術は急速に進歩し、従来は困難であったタスクを実現できるようになってきた。その中でも、特に注目を集めているのが、生成AI、基盤モデル、大規模言語モデルである。下に新技術の特徴を載せる。本報告では大規模言語モデルLLMを中心にスポットを当てる。

技　術	特　徴	応　用　例
生成AI	画像やテキストなどのデータの生成を目的とする	創造的なコンテンツの作成、データの補完
基盤モデル	汎用的なモデルを事前に学習しておくことで、さまざまなタスクに適用可能	自然言語処理、画像認識
大規模言語モデルLLM	言語モデルの一種で、大量のテキストデータを学習することで、人間のような自然な言語を生成できることが特徴	テキスト生成、翻訳、質問応答

表 1: 生成AIのバズワード

1.2　生成AIの課題

生成AIの課題は、生成されたデータの信頼性が挙げられる。生成AIは、データに似た新たなデータを生成することができるが、必ずしも正しいデータとは限らない。そのため、生成されたデータの信頼性を高める技術の開発が重要である。

基盤モデルの課題は、基盤モデルの課題は、学習データの量不足、計算コストの高さ、学習データの偏りの3点である。これらの課題を克服することで、基盤モデルのさらなる発展が期待される。

LLMの課題は、偏りが挙げられる。LLMは膨大な量のテキストデータを学習するため、そのデータに含まれる偏りが、生成されたテキストに反映される可能性がある。そのため、LLMの偏りを軽減する技術の開発が重要である。

2　LLMの基礎技術

2.1　Transformer、Attention、BERTの関係

LLMの基盤技術は「Transformer」と「自己教師あり学習」である。Transformerはディープラーニングの一技術で、従来の方法とは異なる特徴を持つ。従来は各層で入力データの要素が独立して処理されたが、Transformerでは要素間の相互関係を考慮して処理する。この特性により、より複雑なデータに対応可能である。

Transformerの核心は「Attention」メカニズムで、入力データの要素間の関連度をスコアとして計算する。このスコアに基づき、各要素の重要性を判断し、処理を行う。例えば、文書の意味を理解する際、Transformerは各単語の関連度を計算し、重要な単語を特定する。

TransformerとAttentionの関係を簡単に説明すると、TransformerはAttentionを活用して入力と出力の関連性を学習する。Attentionは「Encoder」と「Decoder」の2つの部分から成る。Encoderは入力を特徴量に変換し、Decoderはその特徴量を基に出力を生成する。例として、翻訳タスクでは、Encoderが日本語の文を特徴量に変換し、Decoderがそれを基に英語の文を生成する。

Transformerは、RNNやCNNよりも効率的に関連性を学習することができ、また並列処理が可能になる。BERTはTransformerを基にした言語モデルで、Encoderのみを使用する。BERTはテキストの意味や文脈を捉える特徴量を学習し、分類や要約などのタスクに適用される。BERTはTransformerの能力を最大限に活用し、機械翻訳やテキスト生成、質問応答などのタスクで高い性能を発揮する

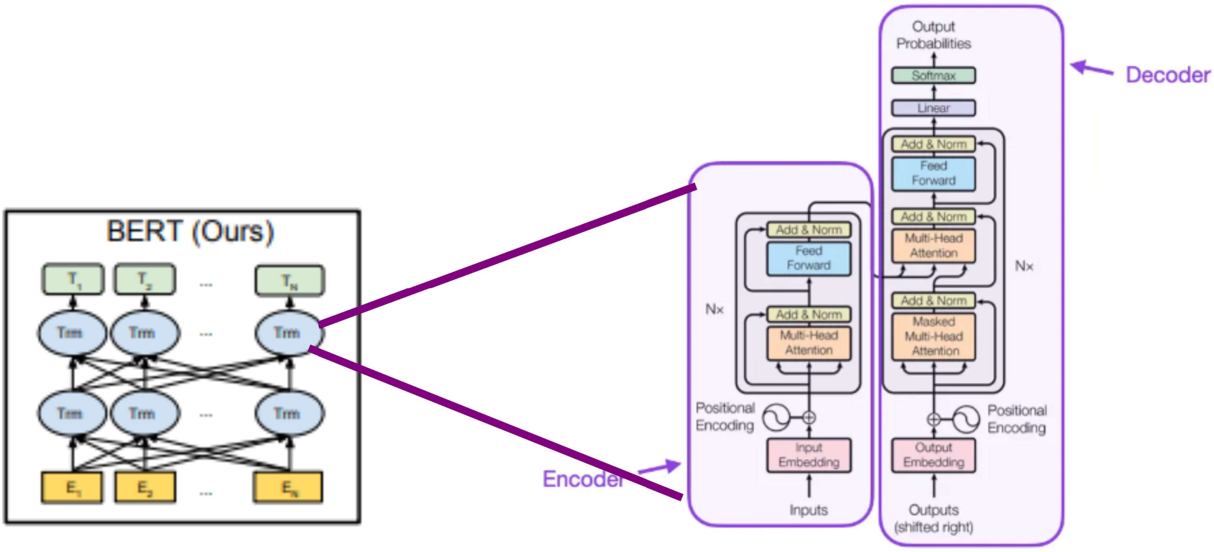


図 1: Attention 機構と BERT の関係図

[Jacob Devlin,2018]

2.2　　自己教師あり学習

最初に自己教師あり学習とほかの学習と合わせて特徴をまとめる。教師あり学習は機械学習でよく使われる自術である。自己教師あり学習は教師あり学習のフレームワークを使用するが、ラベルデータを人手で用意する必要がないという点で、教師なし学習との関連性を有する。

学習方法	特　徴	具　体　例
教師あり学習	教師データとして正解ラベルが与えられている	画像分類、自然言語処理、音声認識、機械翻訳など
教師なし学習	教師データとして正解ラベルが与えられていないため、データの特徴を自動的に学習する	クラスタリング、次元削減、特徴抽出など
自己教師あり学習	教師データとして、元データから何らかの変換や操作を行って「問題」と「答え」のペアを生成し、疑似的なラベルを用いる。このため、人間が明示的にラベルを付ける必要がない	画像生成、テキスト生成、異常検知など

表 2: 機械学習の種類

自己教師あり学習とは、教師データが存在しない場合に、データから学習を行う技術である。LLM は、大量のテキストデータから単語の意味や文の構造を学習する際に、この自己教師あり学習を採用している。自己教師あり学習においては、入力データと出力データの間の対応関係を学習するものである。例として、文書の意味を理解するため、Transformer を活用して、文書の意味を表現するベクトルを出力するモデルの学習が挙げられる。このモデルは、文書の意味を理解するための知識をテキストデータから獲得する。

自己教師あり学習の利点としては、正解データがない場合でも学習が可能であること、学習データ

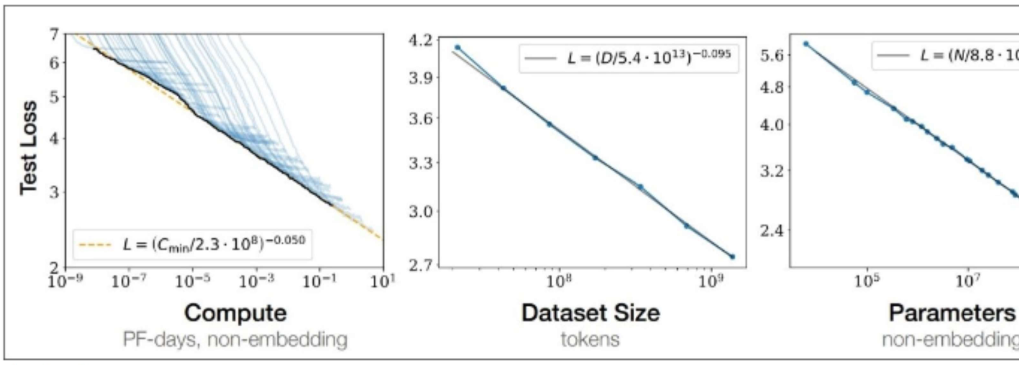
が限られている場合でも効果的な学習が可能であること、学習データの偏りを抑制できることが挙げられる。LLM は、大量のテキストデータを学習の材料とするが、これらのテキストデータが正解データであるとは限らない。しかし、自己教師あり学習の採用により、LLM は正解データが不足していても効果的な学習が行える。

LLM は、Transformer と自己教師あり学習の技術を組み合わせることで、人間に近い自然言語処理の実現を果たした。今後、LLM の技術はさらに進化し、我々の生活に多大な影響をもたらすことが予想される。自己教師あり学習は、LLM の開発における鍵となる技術である。

2.3　パラメーターのサイズと性能の関係

Transformer と自己教師あり学習の組み合わせにより、従来の機械学習では実現できなかった精度が得られた。特に、パラメーターのサイズが大きいモデルでは、より高い精度を実現できるという傾向がある。これは、Transformer では、パラメーターのサイズが大きくなるほど、長距離の依存関係を捉えることが可能になるためである。このため、世界中で巨大なモデル開発競争が勃発した。直近では、中国で 1 兆 7,500 億パラメーターのモデルが開発された。

しかしこの現象は、従来のパラメータサイズの法則が Transformer では破綻することを意味する。統計学や従来の機械学習ではパラメーターの数を大きくするとオーバーフィットするが、Transformer では逆に Test Loss が下がる。理由の詳細はまだわかっていない。[Jared Kaplan,2020]



出所）「Scaling Laws for Neural Language Models(2020)」
左から「計算量」「データ量」「モデルパラメータ数」のべき乗則を示している。

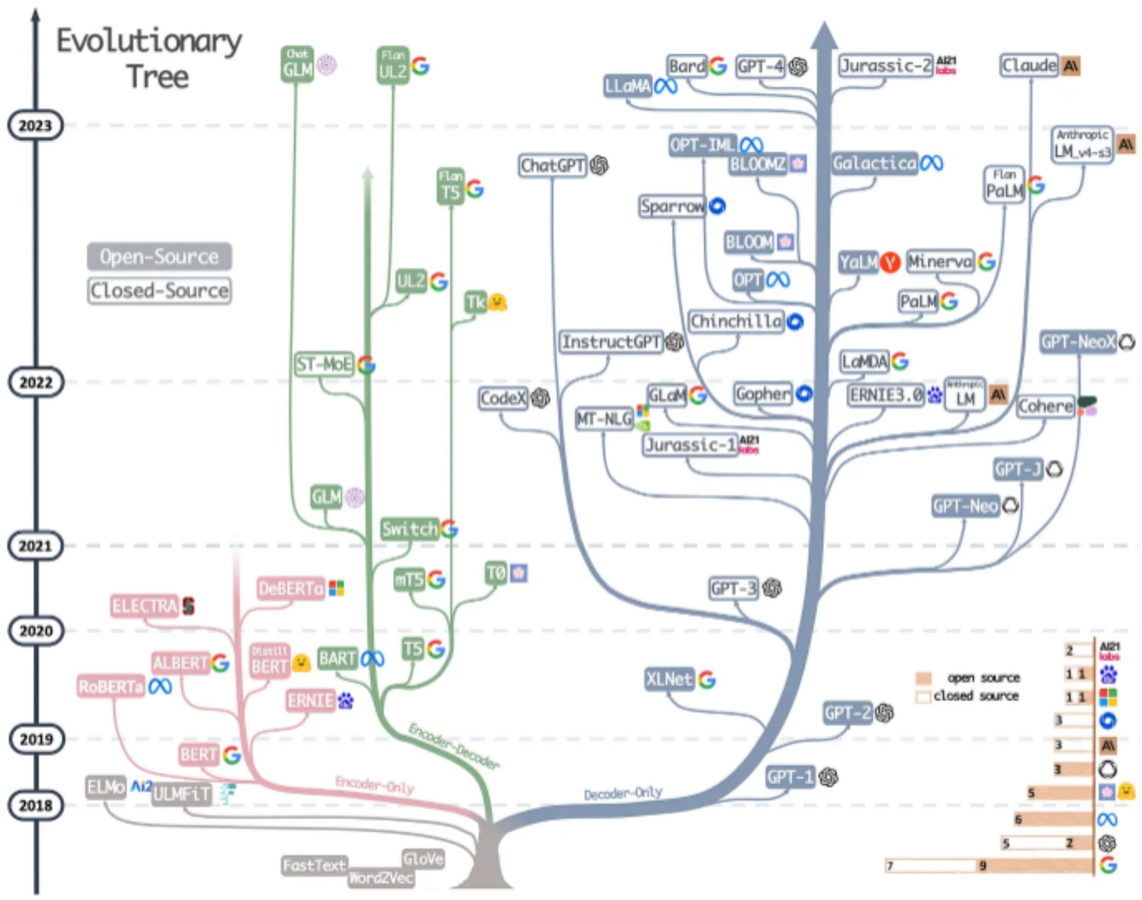
図 2: LLM のパラメーターとサイズの関係

開発会社	開発年度	名　称	パラメータ数	特　徴
OpenAI	2020年	GPT-3	1750億	テキスト生成、翻訳、質問応答などの自然言語処理タスクで高い性能を発揮
OpenAI	2022年	GPT-4	1750億	GPT-3よりも効率的な学習アルゴリズムを採用し、パフォーマンスの向上を実現
Google	2022年	PaLM	5400億	テキスト生成、翻訳、質問応答などの自然言語処理タスクに加え、コードの生成やゲームのプレイなど、幅広いタスクで高い性能を発揮
Microsoft	2021年	Megatron -Turing NLG	5300億	大規模なTransformerモデルで、テキスト生成、翻訳、質問応答などの自然言語処理タスクで高い性能を発揮
AI21 Studio	2023年	Jurassic-1 Jumbo	1.76兆	世界最大級のLLMで、テキスト生成、翻訳、質問応答などの自然言語処理タスクに加え、機械翻訳や要約などのタスクにも活用されている

表 3: LLM のパラメータ数

2.4　　LLM のアーキテクチャの系統

2018 年頃から多種類の LLM が開発された。LLM のアーキテクチャは大きく Encoder-Only, Encoder-Decoder, Decoder-Only の 3 種類に分けられる(図参照)。右の青い枝で表示されているように、特に最近はデコーダー型の発展が顕著であることが見て取れる。[Sebastian Raschka,2023]



https://arxiv.org/abs/2304.13712 を介した現代のLLMの進化ツリー。

図 3: LLM の系列図

3　ChatGPT の登場

3.1　ChatGPT の学習方法

2022 年 11 月 30 日、OpenAI は LLM の一種である ChatGPT を公開した。ChatGPT は、公開後わずか 1 か月で 1 億ユーザーを突破するなど、大きな話題となった。LLM は、膨大な量のテキストデータで事前学習を行うことで、テキストの生成、翻訳、要約、質問応答などのタスクで人間に近い性能を発揮することが可能である。ChatGPT は、GPT-3 をベースにしており、人間との会話を通じて学習する機能を有している。

ChatGPT の学習方法は、大きく分けて 3 つのステップからなる。

第 1 ステップの教師あり学習では、ChatGPT の事前学習モデルに、テキストと対応するテキストのペアを教師データとして与える。教師データは、Web 上から収集したテキストや、人間が生成したテキストなどから構成される。ChatGPT は、この教師データから、テキストの生成に関する一般的な知識を学習する。

第 2 ステップの報酬モデルの学習では、ChatGPT の事前学習モデルを人間と対話させ、その応答を人間が評価する。人間は、ChatGPT の応答が適切かどうか、あるいは誹謗中傷や攻撃的なものになっていないかを判断する。この評価結果は、ChatGPT の学習にフィードバックされる。

第 3 ステップの強化学習では、人間の評価結果に基づいて、ChatGPT の応答を強化学習する。強化学習とは、行動の報酬を最大化するように行動を学習する手法である。ChatGPT は、人間の評価結果を報酬として捉え、その報酬を最大化するように応答を学習する。これはヒューマンフィードバックに基づく強化学習で、RLHF（Reinforcement Learning from Human Feedback）と呼ばれる。

この学習方法により、ChatGPT は人間との会話を通じて、より適切な応答を生成できるようになる。また、誹謗中傷や攻撃的なものにならないよう、適切な行動を学習することもできる。ChatGPT は、強化学習によって炎上しにくいように訓練されているため、多くの人に受け入れられやすい特徴がある。

しかしながら ChatGPT の学習方法には複数の課題が存在する。人間の評価には主観が介入するため、その結果が正確でない可能性がある。また、大量のデータの評価は人間にとって困難である。これらの問題を克服するためには、客観的な評価方法や自動評価システムの導入が求められる。

ChatGPT は、偏見や差別的な表現の生成、誤った情報の生成、そして大量のデータと計算リソースの要求という課題を抱えている。これらの問題を解消するためには、データセットの偏りを排除する手法や、誤情報を検出するアルゴリズムの開発が必要である。さらに、大量のデータ処理を効率的に実行する技術も要される。

4　各クラウド会社の LLM 戦略

4.1　各社の動き

近年、対話型 AI は IT 業界の主要な話題である。大手クラウド企業は、この技術の開発に注力しており、技術の進化に貢献している。

Google は、機械学習の分野でのリーダー企業である。2011 年頃から対話型 AI の研究を始め、2018 年 10 月に「BERT」を発表した。BERT は、自然言語処理技術を用いた対話型 AI で、当時の世界最高水準の文章生成能力を持っていた。しかし、BERT の発表後、Google の対話型 AI の開発は停滞した。その理由として、Amazon の「Alexa」や Apple の「Siri」などの競合製品の市場拡大が挙げられる。

マイクロソフトは、対話型 AI の分野での遅れを取り戻すため、2019 年から OpenAI への出資を開始した。OpenAI も機械学習の分野でのリーダー企業で、2022 年 11 月に「ChatGPT」を発表した。ChatGPT は、BERT を超える性能を持つ対話型 AI で、マイクロソフトはこの技術を「Bing」に取り入れた。Bing の ChatGPT は、対話型の新しい検索機能を持ち、情報の出典の明確化も行われ、実用的なシステムとして評価されている。Amazon は、bedrock や titan を導入し、LLM の展開を進めている。

今後の展望として、OpenAI は 2023 年 3 月 14 日に「GPT-4」を発表した。GPT-4 は、更に進化した対話型 AI で、多くの企業がこの技術の採用を検討していると予想される。

5 生成 AI の歴史

5.1 英語 LLM の歴史

年 号	イベント	内 容
2017年	Transformerの登場	Google のVaswaniらによって、Transformerという新しい機械学習モデルが発表された。Transformerは、従来の機械学習モデルよりも効率的に学習を行うことができるため、LLMの開発を可能にした。
2018年	BERTの登場	BERT（Bidirectional Encoder Representations from Transformers）は、Transformerのアーキテクチャをベースにした言語表現モデルで、テキストの前後の文脈を同時に考慮することで、高い性能を達成している。
2019年	GPT-2の登場	OpenAIがGPT-2を発表。このモデルは、その時点での最先端の性能を持ち、その強力さから、当初は完全なモデルの公開が控えられた。
2019年	RoBERTaの登場	Facebook AIがBERTの改良版としてRoBERTaを発表。 RoBERTaは、BERTの学習方法とデータ処理を改良し、さらなる性能向上を達成した。
2019年	T5の登場	GoogleがT5を発表。T5は、すべてのNLPタスクをテキスト変換タスクとして扱う新しいアプローチを採用。
2019年	DistilBERTの登場	Hugging FaceがBERTの軽量版としてDistilBERTを発表。 DistilBERTは、BERTの性能をほぼ維持しつつ、モデルのサイズを半分に削減。
2019年	XLNetの登場	Google/CMUがBERTとTransformer-XLのアイディアを組み合わせたXLNetを発表。 XLNetは、BERTの双方向性とTransformer-XLの再帰性を組み合わせて、さらなる性能向上を目指した。
2020年	GPT-3の登場	OpenAIのGPT-3が登場し、従来のLLMを大きく上回る性能を実現した。
2022年	対話型AIや画像生成AIの急速な発展	Transformerの登場によって、対話型AIや画像生成AIなどの生成系AIが急速に発展した。
2023年	生成系AIの商用化の動き	Amazon BedrockやZoom IQなどの新サービスが登場し、企業が生成系AIを活用したビジネスを展開する可能性が広がっている。

表 4: 生成 AI モデルの歴史

5.2 ChatGPT 出現後の LLM 激動史

年	月	日	企業名	分野	イベント
2022	11	15	Meta	対話型AI	科学知識に強い対話型AIげGaractica」を試験公開するものの、強い批判を受け2日で公開停止
		30	OpenAI	対話型AI	対話型AIサービス「ChatGPT」を発表
2023	1	17	MicroSoftOpenAI	LLM	大規模言語モデル(LLM)などのサービス「Azure Op日1川Service」を正式提供開始
	2	1	OpenAI	対話型AI	有料サービス「ChatGPT Plus」を発表
		6	Google	対話型AI	対話型AIサービス「B d」を発表
		21	MicroSoft	検索	オープンのLLMをベースにした検索エンシン「Bing」と Webブラウサー「Edge」を発表
		24	Amazon Web Services	LLM	オープンソースのLLMを開発するハギングフェイスと提携
	3	1	Meta	LLM	学習済みのLLMである「LLaMA」をオープンソースとして公開
		7	OpenAI	LLM	ChatGPTや会話認識モテル「Whisper」のAPIを公開
		7	SalesForce	LLM	CRM (顧客関係管理)向けLLMの「Einstein GPT」を発表
		14	OpenAI	LLM	マルチモーダルの基盤モテル「GPT-4」を公開。ChatGPTのバックエンドに使用開始
		14	Google	LLM・開発	基盤モデルのAPIサービス「PaLM API」や 生成搭載アプリケーションの開発ツール「Gen App Builder]を発表
		14	Anthropic	LLM	自社開発LLMの「Claude」を公開
		15	Google	SaaS	自社LLMであるPaLMを搭載したSaaSである「Duet AI　for Google Workspace」を発表
		16	MicroSoft	SaaS	GPT・4を搭載したSaaSである「Microsoft 365 Copilot]を発表
		21	Adobe	画像生成	画像の生成川である「Adobe Firefly」を発表
		22	MicroSoft	開発	子会社の米GitHubがGPT-4をベースにしたソフトウェア開発支援サービスの 「G Hub Copilot X」を発表
		23	OpenAI	対話型AI	ChatGPTの機能を外部サービスによって拡張する 「ChatGPTプラグイン」を発表
		28	MicroSoft	Security	GPT-4を搭載したセキュリティーツール「Microsoft Security C叩i」を発表
	4	12	Databrics	LLM	商用利用が可能な学習済みLLMである「Dolly 2.0」を オープンソースとして公開
		13	Amazon Web Services	LLM	自社LLM「Amazon Titan」や 同モデルのAPIサ-ビス「Amazon　Bedrock」を発表
		14	Google	LLM	医療用のLLMである「Med-PaLM 2」を発表 一部のクラウド顧客向けに提供を始めた
		25	Google	Security	セキュリティー用LLMの「Sec-PaLM」を搭載したセキュリティーツールを発表
	5	10	Google	LLM・SaaS	次世代しLMである「PaLM 2」や、生成AIを組み込んだ新検索エンジン、 生成A陸組み込んだ開発ツール「Duet AI for　Google Cloud」を発表
		18	OpenAI	LLM	ChatGPTのiOS向けアプリケーションを発表
		18	Meta	その他	自社開発AIチップ「MTIA」を発表
		23	MicroSoft	LLM・開発	Windows 11に生成A陸搭載した「Windows Copillot」や、ChatGPTとCopilotのプラグイン互換性などを発表
	6	6	Technology Innovation Intitute(アラブ首長国連邦)	LLM	オープンソースの大規模言語モデル「Falcon－40B、7B」をリリースし 機械学習関連のデータ共有サイト「Hugging Face」にてモデルを公開
		22	MicroSoft	LLM	たった13億のパラメーターでGPT-3.5超えのHumanEval50.6%をたたき出す 「phi-1」を発表
		28	Baidu (中国)	LLM	「GPT-3.5を上回る」AIモデル「Ernie 3.5」を発表
	7	7	上海AI研究所	LLM	GPT-4よりも高い性能を発揮できる特定言語特化型の言語モデル「InternLM」
		19	Meta	LLM	商用可能な大規模言語モデル「Llama 2」を無料公開、 MicrosoftやQualcommと協力してスマホやPCへの最適化も
	8	23	OpenAI	LLM	「GPT-3.5 Turbo」のファインチューニング機能をリリース、 用途に合わせた独自のカスタマイズが可能に
		25	Meta	LLM・開発	商用利用可能なコーディング支援AI「Code Llama」をリリース、 Llama 2と同じライセンスで無料公開へ
		28	OpenAI	対話型AI	企業向けAIチャットサービス「ChatGPT Enterprise」の一般提供開始

表 5: LLM の激動史

（引用文献）日経コンピュータ　2023 年 6 月 22 日号（一部改訂）（ 2023 年 6 月以降は筆者による追加）

5.3　日本語 LLM 発展史

年	月	内　容
2022年	7月	ABEJAがGPT-3をヘースに日本語に特化したLLMの-部を公開
	11月	米オープンAIがGPT-3.5を基に開発した対話型AMChatGPT」を公開
	11月	LINEが820億パラメーターのLLM「HyperCLOVA」について説明
2023年	2月	オルツが1600億パラメーターのLLM「LHTM -2」を開発したと発表
	3月	OpenAIが「GPT4」を公開
		ABEJAがLLMを商用化し、クラウドサービス「ABEJA LLM Series」として提供開始
	5月	ソフトバンクの宮川潤-社長が決算発表の場でLLMの開発意向を表明
		自民党が「国内におけるAI開発基盤を育成・強化すべき」との政策提言を発表
		サイバーエージェントが130億パラメーターのLLMを開発したと発表
		NTTの島田明社長が決算発表の場でLLMの開発意向を表明
		サイバーエーシェントが68億パラメーターのLLMを公開
		rinnaが36億パラメーターのLLMを公開
		富士通や理研がスーパーコンピューター「富岳」を使ったLLMの研究開発を開始
	6月	経団連がLLMなと最先端のAI基盤技術を独自開発すべきとの政策提言を発表
		さくらインターネットが経済産業省の認定のもと、3年で130億円を投じてLLM向けデータセンター環境を整備すると発表
	7月	NECが130億パラメーターで標準的なGPU機で動 するLLMを提供すると発表
	8月	ストックマークが14億パラメータの日本語LLMを公開

表 6: 日本語LLM 発展史

（引用文献）　日経コンピュータ　2023 年 8 月 17 日号（一部改訂）（2023 年 8 月は筆者による追加）

企業名	実用化の時期	パラメーター数	想定用途
ABEJA	2023年3月	130億	生成A使った外部向けのDX支援事業
LINE	未定	820億	文書作成、情報収集、顧客サポートなどの業務効率イヒ(詳細は未定)
NEC	2023年7月	130億	顧客ごとのLLMのカスタマイズ、業種・業務特化の生成Aけービス
NTT	2023年度中	70～300億	NTTグループ各社を通じた業種・業務特化の生成AIサービス
オルツ	2023年5月	1600億	自社の議事録作成サービスやコールセンター向けシステム。外部向けにDX支援事業にも提供
サイバーエージェント	2023年5月	130億	自社のデジタル広告事業、生成AIを使った外部向けのDX支援事業
富士通など	2023年3月以降	1000億程度	富士通のAI基盤サービス群「Kozuchi」に実装して企業向けげサービスとして提供。モテル自体も公開予定
ストックマーク	2023年8月	14億	GPT-NeoXをベースとした14億パラメータの日本語LLM。産総研との共同研究。事前学習はCommonCrawデータでなく、ストックマーク所有の独自Web(2023年6月まで）を使用。

表 7: 日本LLM の概要

（引用文献）日経コンピュータ　2023 年 8 月 17 日号（一部改訂）（ストックマークは筆者による追加）

5.4　日本語 LLM 開発の現状

日本語 LLM の学習データは、非営利団体「Common Crawl」がクロールして集めた日本語データと Wikipedia の日本語版データである。Common Crawl はウェブ上のあらゆる文書を有志が収集し、定期的にスナップショットを作成・公開されているデータである。

日本語 LLM 開発では、英語 LLM と比べてデータ量に大きな差がある。OpneAI の GPT-3 が学習に使ったとされるデータ量は 45 テラバイト。対してネット上に公開されている日本語データセットは数十ギガ～数百ギガバイト、LINE が独自開発した LLM「HyperCLOVA」の学習データでもたかだか 1.8 テラバイトで日本版最大である。そのため、インターネットなどで公開されている日本語データセットを使って学習させたモデルに、特定領域の日本語データを追加学習させ、ファインチューニングを施す。これにより比較的少ないパラメーター数でも勝負できる。2023 年 8 月 29 日に、東京大学松尾研究室では、70 億パラメータの商用利用可能な日本語 LLM『ELYZA-japan-Llama-2-7b』が一般公開された。LLAMA2 を日本語のトークナイザを拡張した上でファインチューニングし、レスポンス性能 1.8 倍で、日本語で最高水準をマークした。つい先日、松尾研究室から GPT-NeoX をベースにファインチューニングした『weblab-10b』が公開に続いた。

ただ、日本国内から Attention を利用した新しい事前学習モデルのアーキテクチャを提案し、そのモデルで最高性能を上回ったというニュースはなく、「GPT-NeoX」「LLaMa2」など、その背後には事前学習済みの「Transformer」ベースの言語モデルをすでに海外の組織が開発を行い、その上で日本語で後から事後学習（ファインチューニングや RLHF)を行ったものがほとんどである。これは、計算資源の問題と、Transformer ベースの言語モデルを実装できる人材が日本では非常に少ないという問題によるものと考えられる。

実務の LLM 開発では、Microsoft の「Azure」上に専用環境を用意し「Azure OpenAI Service」に加え、LLM を使ったサービス開発に有効なライブラリー「LangChain」を併せて開発する。また、今後は Meta の OSS の「Llama2」も視野に入れる必要がある。

5.5　LLM の国際競争力の向上

日本の LLM の国際競争力を高めるためには、バーティカルな LLM の開発、大規模言語モデルの技術的なブレークスルー、そして LLM の可能性を広げる取り組みが重要である。バーティカルな LLM とは、特定のタスクに特化した LLM のことである。日本は多くの業界で高度な技術力を持っているため、バーティカルな LLM の開発に焦点を当てることで、国際的な競争力を向上させることが期待される。

LLM の技術を広げる取り組みとしては、LangChain や LlamaIndex の活用が挙げられる。LangChain は、組織内の文書をインデックス化して質問に対して回答する技術である。LlamaIndex は、テキストとコードを一緒にインデックス化して、質問に対して回答する技術である。これらの技術を活用することで、LLM の可能性はさらに広げることができる。

さらに、LangSmith(2023 年 7 月 18 日公開)という技術も注目されている。LangSmith は、LLM の生成能力をさらに向上させる技術である。LangSmith は、LLM にテキストを生成する際の制約条件を与えることで、より自然で人間らしいテキストを生成する。

6　SAS の LLM

6.1　SAS の LLM の現状

大規模言語モデル（LLM）の開発が進む中、SAS は、LLM を活用した自然言語処理（NLP）への取り組みを積極的に進めている。2020 年には、Google AI が開発した LLM の BERT を活用した NLP 機能を発表し、テキストの分類、要約、翻訳、質問応答などの機能を提供している。 BERT は、テキストの特徴を捉えて分類や回答を行うことができるモデルであり、SAS の NLP 機能では、テキストの感情分析や、商品レビューの要約、翻訳、質問への回答などに利用されている。

また、OpenAI が開発した LLM の ChatGPT を活用した SAS プログラム生成の検証も行っている。ChatGPT は、テキストを生成する機能が優れているため、SAS の NLP 対応に役立つ可能性があると考えられている。しかし、ChatGPT は現状では、中上級レベルの SAS プログラム生成において、誤りやとんでもない嘘プログラミングや検討違いの説明が発生することがある。そのため、SAS は ChatGPT の将来性については、まだ懐疑的であるようだ。

SAS は LLM の将来性に期待を寄せているが、LLM の動向を注目している段階で、SAS への取り入れるかどうかの発表はまだない。また、SAS プログラムの Copilot の開発が急募されている。Copilot はプログラム生成の機能を強化したもので、仕様書まで生成することができる。7 月に公開された GitHub 等の Copilot で確認した結果、SAS の仕様や上級プログラムを生成するにはまだ時間が必要であることが判明した。各社の Copilot は Python を中心に多くの言語のプログラムや仕様書を生成する能力があるようだが、SAS にはまだ対応していない。このため、SAS 社自身で Copilot の開発が必要である。

年	イベント・発表内容
2020	SAS GLOBAL FORUM (SAS 社の世界大会、米国開催)にてSAS のBERT を発表
2021	SAS のAntti Heino 氏がYouTube でSAS のBERT の利用方法を2 本発表
2022	SAS Institute Japan の自然言語処理担当の堀内氏がAntti Heino 氏のYouTube を参考にSAS を利用した日本語の自然言語処理をSAS Blog 上で発表

表 8: SAS の LLM に関するイベント

6.2　SAS BERT のリンク先

①　SAS GLOBAL　FORUM　2020

NLP with BERT: Sentiment Analysis Using SAS® Deep Learning and DLPy　 SAS_BERT_4429-2020.pdf

https://support.sas.com/resources/papers/proceedings20/4429-2020.pdf

②　SAS の自然言語処理

・SAS　Visual　Text　Analytics　プログラミング

https://go.documentation.sas.com/doc/en/pgmsascdc/v_040/casvtapg/titlepage.htm

・Text Classifier Action Set : SAS Viya　テキスト分類の構文

Natural Language Processing (NLP) - Score Text Classifier
https://github.com/sassoftware/sas-studio-custom-steps/blob/main/NLP%20-%20Score%20Text%20Classifier/README.md

・scoreTextClassifier Action
https://go.documentation.sas.com/doc/en/pgmsascdc/v_040/casvtapg/p179jrv0ml345an1emuoxwmccm1j.htm#n0qmia42fpwpoxn1dqemoxt34uhu

・SAS Documentation
https://go.documentation.sas.com/doc/en/pgmsascdc/v_040/casvtapg/cas-textclassifier-scoretextclassifier.htm?fromDefault=

③　Youtube　SAS の BERT の使い方解説　（Antti Heino、SAS 社）

・Introduction to BERT and how to implement it in SAS Viya
https://www.youtube.com/watch?v=Wrcx3ZwqaD4

・BERT for Text Classification on SAS Viya
https://www.youtube.com/watch?v=-Xml2k1CYlk

④　SAS の日本語 BERT
堀内氏　（SAS 社自然言語処理担当）
https://blogs.sas.com/content/author/ryosukehoriuchi/2022.09.20

7　終わりに

筆者は LLM の開発に携わっている。この技術は非常に早い進化を遂げており、毎週のように新しい理論や LLM、ツールが発表されている。そのため、3 か月前の知識では、時代遅れになってしまう。

LLM の開発には、膨大な計算リソースと、専門的な知識が必要である。そのため、日本では、米国よりも開発の周遅れが出ている。しかし、日本には、優れたエンジニアや研究者が多く存在する。今後は、これらの人材を結集し、日本語 LLM の開発を加速させることが重要である。また、LLM の可能性を広げる取り組みとして LangChain や LlamaIndex や LangSmith などの新技術の活用することが重要。

また統計ツールとして SAS も、SAS 版 LLM を利用した Copilot の装備が待たれる。

本報告は、2023 年 8 月 30 日までの情報を基に作成した。

8　参考文献

[Jared Kaplan,2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei, "Scaling Laws for Neural Language Models",2020,https://arxiv.org/abs/2001.08361

[JINGFENG YANG,2023]JINGFENG YANG, HONGYE JIN, RUIXIANG TANG, XIAOTIAN HANK, QIZHANG FENG, HAOMING JIANG, BING YIN, XIA HU ,"Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond",2023,https://arxiv.org/abs/2304.13712

[Ashish Vaswani,2017]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin,"Attention Is All You Need",2017,https://arxiv.org/abs/1706.03762

STREAM Procedureを用いたMedical Writingの効率化

○平井 隆幸[1]、今泉 敦[1]、叶 健[1]、金子 幸子[1]、三澤 早織[1]

（[1]日本化薬株式会社）

Streamlining the process of Medical Writing using the STREAM Procedure

Takayuki Hirai[1], Atsushi Imaizumi[1], Takeshi Kanou[1], Sachiko Kaneko[1], and Saori Misawa[1]
Biostatistics Team, Clinical Development Strategy Division, Nippon Kayaku Co.,Ltd

要旨

本稿では、SAS 9.4 から正規版として利用ができるようになった STREAM Procedure を用いて、Writer が表を基にマニュアルで作成していた定型文書を半自動で作成する方法について報告する。

キーワード：STREAM Procedure、マクロ、テキスト、有害事象の集計表、定型の文書、半自動

1. はじめに

医薬品開発では、早期の承認申請を目指して、常に高品質のドキュメントを短期間に作成することが求められる。しかし、製薬企業にとって Medical Writing（MW）部門の人員を増やすことも、一度に多数の有能かつ経験豊富な Medical Writer を育成・確保することも簡単でないのが現状である。こうした背景から、製薬企業では、Medical Writer が執筆自体に注力できるように、「人」による高い専門技術や創造性が求められる仕事とそうでない仕事を分け、後者に対しては積極的にデジタル技術を活用するという取り組みが進められている。たとえば、<u>①これまで Medical Writer が原資料をもとにマニュアルで作成していた定型の文書を、自動で作成するための支援ツールを導入する</u>、あるいは<u>②構造化された文書間で記述を再利用できるシステムを開発する</u>というものである（2019 年、日本メディカルライター協会第 18 回シンポジウム要旨より引用）。

STREAM Procedure とは、マクロによる指定を含むテキストを、マクロから得られる結果と共に外部ファイルに出力することができる Procedure であり、SAS 9.4 から正規版として利用ができるようになった。SAS Viya では、SAS 9.4M5 からサポートが追加されるようになった。HTML、XML、RTF 等のテキストベースファイルに使用され、マクロを実行し展開する間、入力ストリームの他のテキストは保持され、SAS 構文として検証されない特徴を持つ。PROC STREAM を介し実行することで、HTML や XML のタグ、RTF コードによる SAS 構文 Error を回避し、外部ファイルに出力でき、関根（2017）や山崎ら（2018）で利用事例が報告されている。本稿では、上記の①を目的とした支援ツールとして、STREAM Procedure を用いて、Medical Writer が、有害事象の集計表と文書テンプレートを基に、有害事象名や発現例数、発現割合等の内容についてマニュアルで作成していた定型の文書を、半自動で作成することができないか検討を行い、活用した事例を報告する。

2. STREAM Procedure の基本的な使い方

2.1. 基本構文

① 「OUTFILE=file1」は、すべてのトークンが書き込まれるファイルを指定する。PROC STREAM ステートメントでは、「OUTFILE=」キーワード及びオプションを使用して外部ファイルを指定する。

② 入力ストリームは、任意のテキストを「BEGIN」と 4 つのセミコロン末尾「;;;;」で囲む。
STREAM Procedure の開始は、入力ストリームの開始を識別するキーワード「BEGIN」を記述する。
STREAM Procedure の終了は、「RUN;」でも「QUIT;」でもなく、プロシジャの最後の行として、間にブランクを含まない 4 つのセミコロン「;;;;」を記述する。「;;;;」が入力ストリームの終了を識別する。

③ text-1 : PROC STREAM で使用する SAS ステートメントまたはマクロを指定する。

左のコードを実行することで、右の出力が作成できる。詳細は SAS Institute Inc（2023）を参照頂きたい。

```
filename file1 "C:¥temp¥file1.txt";
PROC STREAM OUTFILE=file1;
BEGIN
text-1<text-n>
;;;;
```

⇒

```
text-1<text-n>
```

2.2. 出力ストリームへの新しい行の挿入

入力ストリームにおいて、RESETDELIM=オプションを使用して指定したデリミタ「goto」の後ろのセミコロン「;」の前にキーワード「NEWLINE」を追加することで、出力ストリームへの新しい行の挿入が可能である。

```
filename file2 "C:¥temp¥file2.txt";
PROC STREAM OUTFILE=file2 RESETDELIM="goto";
BEGIN
text-1 goto NEWLINE;
<text-n>
;;;;
```

⇒

```
text-1
<text-n>
```

2.3. %INCLUDE を使用した PROC STREAM へのファイルの挿入

入力ストリームにおいて、%INCLUDE や%LET ステートメントを展開する場合は、RESETDELIM=オプションを使用して指定したデリミタ「goto」とセミコロン「;」を%INCLUDE や%LET ステートメントの前に挿入する。これは、%INCLUDE や%LET ステートメントは、ステートメント境界から開始する必要があるためである。%INCLUDE を使用した PROC STREAM へのファイルの挿入方法を示す。

```
filename file3 "C:¥temp¥file3.txt";
PROC STREAM OUTFILE=file3 RESETDELIM="goto";
BEGIN
text-0 goto; /*%INCLUDEの前、即ちtext-0の後にgoto;*/
%INCLUDE file1; /*%INCLUDEでfile1を展開*/
;;;;
```

⇒

```
text-0text-1<text-n>
```

3. マクロによる指定を含むテキストを外部ファイルに出力

3.1. シンプルなマクロベースコードの展開

STREAM Procedure は、入力ストリームにおいて、マクロによる指定を含むテキストを入力し、実行することで、マクロから得られる結果と共に外部ファイルに出力することができる。

```
%Let TableT=有害事象の一覧;%Let TableN=表12;
filename file4 "C:¥temp¥file4.txt";
PROC STREAM OUTFILE=file4 RESETDELIM="goto";
BEGIN goto;
&TableT.を&TableN.に示した。
;;;;
```

⇒

有害事象の一覧を表12に示した。

3.2. データに基づく判断を加えたマクロベースコードの展開

入力ストリームにおいて、有害事象の発現者数を表す変数 Gnp（Number & Percent of patients in Group）のデータから文中で用いる文書テンプレートを判断し挿入するマクロ%Tmp を含むテキストを入力し、実行することで、データに基づく判断を加えたマクロから得られる結果と共に外部ファイルに出力することもできる。

```
/*データに基づく判断を加えたマクロ*/
%macro Tmp();
%if "&Gnp." = "0例" %then %do;
有害事象は&Gnp.であり、認められなかった。
%end;
%else %if "&Gnp." ^="0例" %then %do;
有害事象は&Gnp.に認められた。
%end;
%mend Tmp;
/*有害事象の発現者数0例の場合*/
%Let Gnp=0例;
filename file5 "C:¥temp¥file5.txt";
PROC STREAM OUTFILE=file5 RESETDELIM="goto";
BEGIN goto;
&TableT.を&TableN.に示した。%Tmp()
;;;;
```

⇒

有害事象の一覧を表12に示した。有害事象は0例であり、認められなかった。

```
/*有害事象の発現者数0例でない場合*/
%Let Gnp=50/100例（50.0%NRBQUOTE(%)）;
PROC STREAM OUTFILE=file5 RESETDELIM="goto";
BEGIN goto;
&TableT.を&TableN.に示した。%Tmp()
;;;;
```

⇒

有害事象の一覧を表12に示した。有害事象は50/100例（50.0%）に認められた。

4.　Medical Writing への適用

4.1.　検討事例

Data Science/Biostatistics（DS/BS）部門で Clinical Study Report（CSR）、Common Technical Document（CTD）、照会事項回答用に作成した有害事象の集計表と MW 部門で保有する CSR/CTD/照会事項回答用の文書テンプレートを基に、Medical Writer がマニュアルで定型の文書を作成する。具体的には、下記の有害事象の集計表に対する文書テンプレートが、●を用いて表記されており、集計表からマニュアルで群ごとに●を埋める形で定型の文書を作成する。有害事象の集計表と文書テンプレートを基に Medical Writer がマニュアルで作成するこの定型の文書を、STREAM Procedure を用いて、半自動で作成することができないか検討を行った。

Table X.X　Summary of TEAEs by PT／Safety Population（X Study）

PT Name	Treatment N=xxx n (%)	Placebo N=xxx n (%)
Subjects with Any TEAEs	xxx (xxx.x)	xxx (xxx.x)
PT Name#1	xxx (xxx.x)	xxx (xxx.x)
PT Name#2	xxx (xxx.x)	xxx (xxx.x)

MedDRA Version XX.X.

有害事象は、●/●例（●%）に認められた。認められた有害事象は、●●　●例（●%）であった。

4.2.　検討に用いたデータセット

吉田（2017）で用いたデータセット_ALL_TEAEs_F を使い、有害事象の集計表と文書テンプレートを基にマニュアルで作成していた定型の文書を、STREAM Procedure を用いて、半自動で作成する方法を説明する。日本語の文書テンプレートを用いるため、日本語の有害事象名を表す変数 JPT を追加し、PT Name 列に用いる。

ITEM_NO	ITEM_DISP	JPT	ABC_XXX	Placebo	indentwt
0	Subjects with Any TEAEs	有害事象の発現者数	43 (43.0)	42 (42.0)	0
1	Abdominal Pain	腹痛	4 (4.0)	6 (6.0)	1
2	Arrhythmia	不整脈	6 (6.0)	2 (2.0)	1
3	Constipation	便秘	7 (7.0)	4 (4.0)	1
4	Dizziness	浮動性めまい	3 (3.0)	3 (3.0)	1
5	Headache	頭痛	6 (6.0)	7 (7.0)	1
6	Influenza	インフルエンザ	4 (4.0)	2 (2.0)	1
7	Liver Function Test Abnormal	肝機能検査異常	3 (3.0)	6 (6.0)	1
8	Nausea	悪心	4 (4.0)	4 (4.0)	1
9	Swelling Face	顔面腫脹	5 (5.0)	6 (6.0)	1
10	Upper Respiratory Tract Infection	上気道感染	4 (4.0)	2 (2.0)	1
11	Vomiting	嘔吐	5 (5.0)	6 (6.0)	1

4.3.　Medical Writer がマニュアルで作成する場合

4.2 節のデータセットを用いて作成した有害事象の集計表と文書テンプレートを基に、マニュアルで定型の文書を作成すると、次のようになる。4.1 節で示した文書テンプレート中の●は、各群表記、各群の有害事象の発現例数、安全性解析対象集団の例数、発現割合（%）、各群の発現例数の多い有害事象から順に有害事象名、各有害事象の発現例数及び発現割合（%）の一覧をそれぞれ表す。単純に有害事象の集計表の内容を定型の

文書に書き起こすだけであるが、各群の発現例数の多い有害事象から順にマニュアルで記載し、記載が正しいかどうかを集計表と突合せた後、記載を固定する作業は、作成と QC に多くの工数を要する。抗がん剤等、領域によっては発現する有害事象の数が多く、大変な作業になる。また、片方の群のみ有害事象が発生する場合や両群ともに発生しない場合、データの状況に応じて、文書テンプレートの●以外を書き換える必要がある。●以外の書き換えを各 Medical Writer が行うことで、表記ゆれが生じ、MW の品質低下に繋がる可能性がある。

表 X.X　有害事象の発現者数／安全性解析対象集団（X 試験）

基本語	ABC_XXX N=100 n (%)	Placebo N=100 n (%)
有害事象の発現者数	43 (43.0)	42 (42.0)
腹痛	4 (4.0)	6 (6.0)
不整脈	6 (6.0)	2 (2.0)
便秘	7 (7.0)	4 (4.0)
浮動性めまい	3 (3.0)	3 (3.0)
頭痛	6 (6.0)	7 (7.0)
インフルエンザ	4 (4.0)	2 (2.0)
肝機能検査異常	3 (3.0)	6 (6.0)
悪心	4 (4.0)	4 (4.0)
顔面腫脹	5 (5.0)	6 (6.0)
上気道感染	4 (4.0)	2 (2.0)
嘔吐	5 (5.0)	6 (6.0)

MedDRA Version XX.X.

有害事象は、**ABC_XXX 群 43/100 例（43.0%）**、**プラセボ群 42/100 例（42.0%）**に認められた。
ABC_XXX 群で認められた有害事象は、**便秘 7 例（7.0%）、不整脈及び頭痛　各 6 例（6.0%）、顔面腫脹及び嘔吐　各 5 例（5.0%）、腹痛、インフルエンザ、悪心及び上気道感染　各 4 例（4.0%）、浮動性めまい及び肝機能検査異常　各 3 例（3.0%）**であった。
プラセボ群で認められた有害事象は、**頭痛　7 例（7.0%）、腹痛、肝機能検査異常、顔面腫脹及び嘔吐　各 6 例（6.0%）、便秘及び悪心　各 4 例（4.0%）、浮動性めまい 3 例（3.0%）、不整脈、インフルエンザ及び上気道感染　各 2 例（2.0%）**であった。

4.4.　STREAM Procedure を用いて作成する場合

次に STREAM Procedure を用いて、有害事象の集計表と文書テンプレートを基に、定型の文書を半自動で作成する方法を説明する。

① 文書テンプレートを STREAM Procedure の入力ストリームに当てはめ、ベースとなる記載を作成する。集計表から Medical Writer が埋めていた●をマクロ変数 **G1/G2（Group i）**、**Gnp1/Gnp2（Number & Percent of patients in Group i）**、**JPTList1/JPTList2（Japanese Preferred Term List i）**で定義して記載する。

有害事象は、&G1.&Gnp1.、&G2.&Gnp2.に認められた。&G1.で認められた有害事象は、&JPTList1.であった。&G2.で認められた有害事象は、&JPTList2.であった。

② データの状況に基づき、要求される記載を条件分岐させる。各記載は、MW 部門と協議し一貫性のある記載を作成することを推奨する。下記を、STREAM Procedure の入力ストリームに挿入することで、文書

テンプレートの●以外の記載は、全て表記ゆれなく、自動で作成が可能になる。

G1 の AE 有無	G2 の AE 有無	データの状況に基づき条件分岐した記載
1	1	有害事象は、&G1.&Gnp1.、&G2.&Gnp2.に認められた。&G1.で認められた有害事象は、&JPTList1.であった。&G2.で認められた有害事象は、&JPTList2.であった。
1	0	有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、&G1.のみに認められた。&G1.で認められた有害事象は、&JPTList1.であった。
0	1	有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、&G2.のみに認められた。&G2.で認められた有害事象は、&JPTList2.であった。
0	0	有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、認められなかった。

③ 群表記のマクロ変数 G1 と G2 を指定する。集計表作成時に指定があり、文書作成時の群表記と同じ場合は不要で、集計表作成時の群表記を再利用する。文書作成時の群表記と異なる場合は、STREAM Procedure の実行前に、%LET ステートメントを用いて再指定することで群表記を変更する。提出する Documents（CTD や照会事項回答等）においては、集計表作成時とは異なる群表記を要求される場合もある。例えば、新薬の場合：試験治療群と標準治療群、バイオシミラー（BS）の場合：本剤群と先行バイオ医薬品群、ジェネリック医薬品（GE）の場合：試験製剤群と標準製剤群等の表記を要求される。集計表作成時の群表記指定から再指定し集計表から再作成する、または文書作成時の群表記を変更する等で対応する。

集計表

基本語	ABC_XXX N=100 n (%)	Placebo N=100 n (%)
有害事象の発現者数	43 (43.0)	42 (42.0)
腹痛	4 (4.0)	6 (6.0)
不整脈	6 (6.0)	2 (2.0)
便秘	7 (7.0)	4 (4.0)
浮動性めまい	3 (3.0)	3 (3.0)
頭痛	6 (6.0)	7 (7.0)
インフルエンザ	4 (4.0)	2 (2.0)
肝機能検査異常	3 (3.0)	6 (6.0)
悪心	4 (4.0)	4 (4.0)
顔面腫脹	5 (5.0)	6 (6.0)
上気道感染	4 (4.0)	2 (2.0)
嘔吐	5 (5.0)	6 (6.0)

STREAM Procedureの入/出力

【入力ストリームに挿入する文書テンプレート】
有害事象は、**&G1.**&Gnp1.、**&G2.**&Gnp2.に認められた。**&G1.**で認められた有害事象は、&JPTList1.であった。**&G2.**で認められた有害事象は、&JPTList2.であった。

【出力ストリームの結果】
有害事象は、**ABC_XXX群**43/100例（43.0%）、**プラセボ群**42/100例（42.0%）に認められた。
ABC_XXX群で認められた有害事象・・・であった。
プラセボ群で認められた有害事象・・・であった。

マクロ変数G1とG2作成過程コードとイメージ

%let G1=ABC_XXX群; %let G2=プラセボ群;

集計表作成時の群表記指定（例）		文書作成時の群表記指定（例）
%let G1=ABC_XXX; %let G2=Placebo;	⇒	%let G1=ABC_XXX群; %let G2=プラセボ群;
%let G1=ABC_XXX群; %let G2=プラセボ群;	⇒	指定不要（※集計表作成時の群表記を再利用）
%let G1=Drug A; %let G2= Drug B;（新薬の場合）	⇒	%let G1=試験治療群; %let G2=標準治療群;
%let G1=Drug A; %let G2= Drug B;（BSの場合）	⇒	%let G1=本剤群; %let G2=先行バイオ医薬品群;
%let G1=Drug A; %let G2= Drug B;（GEの場合）	⇒	%let G1=試験製剤群; %let G2=標準製剤群;

④　各群の有害事象の発現例数、安全性解析対象集団の例数、発現割合（%）を表記するマクロ変数 Gnp1 と Gnp2 は、次の通り、集計表作成時のデータセットを再利用し指定する。

集計表

基本語	ABC_XXX N=100 n (%)	Placebo N=100 n (%)
有害事象の発現者数	43 (43.0)	42 (42.0)
腹痛	4 (4.0)	6 (6.0)
不整脈	6 (6.0)	2 (2.0)
便秘	7 (7.0)	4 (4.0)
浮動性めまい	3 (3.0)	3 (3.0)
頭痛	6 (6.0)	7 (7.0)
インフルエンザ	4 (4.0)	2 (2.0)
肝機能検査異常	3 (3.0)	6 (6.0)
悪心	4 (4.0)	4 (4.0)
顔面腫脹	5 (5.0)	6 (6.0)
上気道感染	4 (4.0)	2 (2.0)
嘔吐	5 (5.0)	6 (6.0)

STREAM Procedureの入/出力

【入力ストリームに挿入する文書テンプレート】

有害事象は、&G1.**&Gnp1.**、&G2.**&Gnp2.**に認められた。・・・

↓

【出力ストリームの結果】

有害事象は、ABC_XXX群**43/100例（43.0%）**、プラセボ群**42/100例（42.0%）**に認められた。・・・

マクロ変数Gnp1とGnp2作成過程コードとイメージ

```
/*各群の安全性解析対象集団の例数をSAFN1とSAFN2に指定（集計表作成時に指定済の場合不要）*/
proc sql noprint ;
select count(distinct USUBJID) into :SAFN1-:SAFN2 from ADSL where SAFFL = "Y" group by TRT01AN;
quit ;
/*1.集計表を出力する直前のDatasetをsetし、処理変数にrename*/
data OUTPUT;
set _ALL_TEAEs_Fj;
rename JPT=Title    ABC_XXX=pcol1    Placebo=pcol2;
run;
/*2.集計表のTEAEの発現者数の行のDataを抽出*/
data OUTPUT_D;
length pcol1 $32767. pcol2 $32767.;
set OUTPUT;
where Title="有害事象の発現者数";
/*3.TRANWRD関数を用いて半角「(」を「/N例（」へ置換*/
/*TRANWRD(対象変数,置換したい文字列,置換後文字列)*/
pcol1=tranwrd(pcol1, " (", "/&SAFN1.例（");    pcol2=tranwrd(pcol2, " (", "/&SAFN2.例（");
/*発現者数0の場合は0例へ置換*/
if pcol1="0" then    pcol1="0例";    if pcol2="0" then    pcol2="0例";
run;


/*4.再度TRANWRD関数を用いて半角「)」を「%NRBQUOTE(%))」へ置換*/
/*%は記述文字とするため、%NRBQUOTE関数を用いてマスク*/
```

Step.1

title	pcol1	pcol2
有害事象の発現者数	43 (43.0)	42 (42.0)
腹痛	4 (4.0)	6 (6.0)
不整脈	6 (6.0)	2 (2.0)
便秘	7 (7.0)	4 (4.0)
浮動性めまい	3 (3.0)	3 (3.0)
頭痛	6 (6.0)	7 (7.0)
インフルエンザ	4 (4.0)	2 (2.0)
肝機能検査異常	3 (3.0)	6 (6.0)
悪心	4 (4.0)	4 (4.0)
顔面腫脹	5 (5.0)	6 (6.0)
上気道感染	4 (4.0)	2 (2.0)
嘔吐	5 (5.0)	6 (6.0)

Step.2

pcol1	pcol2
43 (43.0)	42 (42.0)

Step.3

pcol1	pcol2
43/100例(43.0)	42/100例(42.0)

Step.4

pcol1	pcol2
43/100例(43.0%NRBQUOTE(%))	42/100例(42.0%NRBQUOTE(%))

```
data OUTPUT_D;
set OUTPUT_D;
pcol1=tranwrd(pcol1, ")", "%NRBQUOTE(%)）");    pcol2=tranwrd(pcol2, ")", "%NRBQUOTE(%)）");
run;
/*5.記載情報を変数pcol1とpcol2に要約しマクロ変数Gnp1とGnp2に格納*/
data _NULL_;
set OUTPUT_D;
call symputx("Gnp1",pcol1);    call symputx("Gnp2",pcol2);
run;
```

⑤　各群の発現例数の多い有害事象から順に有害事象名、各有害事象の発現例数及び発現割合（%）の一覧を表記するマクロ変数 JPTList1 と JPTList2 も、集計表作成時のデータセットを再利用し指定する。

集計表

基本語	ABC_XXX N=100 n (%)	Placebo N=100 n (%)
有害事象の発現者数	43 (43.0)	42 (42.0)
腹痛	4 (4.0)	6 (6.0)
不整脈	6 (6.0)	2 (2.0)
便秘	7 (7.0)	4 (4.0)
浮動性めまい	3 (3.0)	3 (3.0)
頭痛	6 (6.0)	7 (7.0)
インフルエンザ	4 (4.0)	2 (2.0)
肝機能検査異常	3 (3.0)	6 (6.0)
悪心	4 (4.0)	4 (4.0)
顔面腫脹	5 (5.0)	6 (6.0)
上気道感染	4 (4.0)	2 (2.0)
嘔吐	5 (5.0)	6 (6.0)

STREAM Procedureの入/出力

【入力ストリームに挿入する文書テンプレート】

&G1.で認められた有害事象は、**&JPTList1.**であった。

&G2.で認められた有害事象は、**&JPTList2.**であった。

↓

【出力ストリームの結果】

ABC_XXX群で認められた有害事象は、**便秘7例（7.0%）、不整脈及び頭痛 各6例（6.0%）、顔面腫脹及び嘔吐 各5例（5.0%）、腹痛、インフルエンザ、悪心及び上気道感染 各4例（4.0%）、浮動性めまい及び肝機能検査異常 各3例（3.0%）**であった。

プラセボ群で認められた有害事象は、**頭痛7例（7.0%）、腹痛、肝機能検査異常、顔面腫脹及び嘔吐 各6例（6.0%）、便秘及び悪心 各4例（4.0%）、浮動性めまい3例（3.0%）、不整脈、インフルエンザ及び上気道感染 各2例（2.0%）**であった。

マクロ変数JPTList1作成過程コードとイメージ（JPTList2も同様に作成）

```
/*1.発現PTのみ抽出し、pcol1のn数を数値変数pt_1nと文字変数pcol1nへ、Percentを文字変数pcol1pへ格納*/
data OUTPUT_D;
set OUTPUT;
where indentwt=1 and pcol1^='0'; /*発現PTのみ抽出*/
pt_1n=input(SCAN(pcol1,1,'('),best12.);
pcol1n=SCAN(pcol1,1,'(');
pcol1p=SCAN(SCAN(pcol1,2,'('),1,')');
run;
```

Step.1

title	pcol1	pt_1n	pcol1n	pcol1p
腹痛	4 (4.0)	4	4	4.0
不整脈	6 (6.0)	6	6	6.0
便秘	7 (7.0)	7	7	7.0
浮動性めまい	3 (3.0)	3	3	3.0
頭痛	6 (6.0)	6	6	6.0
インフルエンザ	4 (4.0)	4	4	4.0
肝機能検査異常	3 (3.0)	3	3	3.0
悪心	4 (4.0)	4	4	4.0
顔面腫脹	5 (5.0)	5	5	5.0
上気道感染	4 (4.0)	4	4	4.0
嘔吐	5 (5.0)	5	5	5.0

```
/*2.出力順は各群の発現例数の多いPT順であるため、descending pt_1nでsortする*/
proc sort data=OUTPUT_D;
by descending pt_1n;
run;
```

Step.2

title	pcol1	pt_1n	pcol1n	pcol1p
便秘	7 (7.0)	7	7	7.0
不整脈	6 (6.0)	6	6	6.0
頭痛	6 (6.0)	6	6	6.0
顔面腫脹	5 (5.0)	5	5	5.0
嘔吐	5 (5.0)	5	5	5.0
腹痛	4 (4.0)	4	4	4.0
インフルエンザ	4 (4.0)	4	4	4.0
悪心	4 (4.0)	4	4	4.0
上気道感染	4 (4.0)	4	4	4.0
浮動性めまい	3 (3.0)	3	3	3.0
肝機能検査異常	3 (3.0)	3	3	3.0

```
/*3.発現例数別のPT Listを表す新規文字変数PTList1を作成する*/
data OUTPUT_D;
  set OUTPUT_D;
by descending pt_1n;
length PTList1 $32767.;
*  新規文字変数PTList1に値を保持する;
  retain PTList1;
* pt_1nすなわち発現例数の多いPTから順に、1Observation（Obs）目の時、Titleの値をPTList1に格納;
  if first.pt_1n then PTList1=Title;
*1Obs目でないかつ最後のObs目でない時、PTList1の値とTitleの値を「、」で結合してPTList1に格納;
  else if last.pt_1n^=1 then    PTList1=catx('、',PTList1,Title);
*1Obs目でないかつ最後のObs目である時、PTList1の値とTitleの値を「及び」で結合してPTList1に格納;
  else    PTList1=catx('及び',PTList1,Title);
*そのpt_1nの最後のObsの時（例：pt_1n=6の時は頭痛）、outputした後、PTList1を初期化（欠損値に）する;
  if last.pt_1n then do;
    output;
    call missing(of PTList1);
  end;
run;
```

Step.3

title	pcol1	pt_1n	pcol1	pcol1p	PTList1
便秘	7 (7.0)	7	7	7.0	便秘
頭痛	6 (6.0)	6	6	6.0	不整脈及び頭痛
嘔吐	5 (5.0)	5	5	5.0	顔面腫脹及び嘔吐
上気道感染	4 (4.0)	4	4	4.0	腹痛、インフルエンザ、悪心及び上気道感染
肝機能検査異常	3 (3.0)	3	3	3.0	浮動性めまい及び肝機能検査異常

```
/*4.PTList1に発現例数と発現割合を付加した新規文字変数NP_PTList1を作成する*/
data OUTPUT_D;
length NP_PTList1 $32767.;
set OUTPUT_D;
*TitleとPTList1が同じ場合は、1つのPTに対する記載としてNP_PTList1を出力する;
if Title=compress(PTList1) then NP_PTList1=compress(PTList1) || ' ' || compress(pcol1n) || '例（' || compress(pcol1p)
|| '%NRBQUOTE(%))';
*TitleとPTList1が異なる場合は、複数のPTに対する記載としてNP_PTList1を出力する;
else NP_PTList1=compress(PTList1) || ' 各' || compress(pcol1n) || '例（' || compress(pcol1p) || '%NRBQUOTE(%))
';
run;
```

Step.4

title	pcol1	pt_1n	pcol1n	pcol1p	NP_PTList1
便秘	7 (7.0)	7	7	7.0	便秘 7例(7.0%NRBQUOTE(%))
頭痛	6 (6.0)	6	6	6.0	不整脈及び頭痛 各6例(6.0%NRBQUOTE(%))
嘔吐	5 (5.0)	5	5	5.0	顔面腫脹及び嘔吐 各5例(5.0%NRBQUOTE(%))
上気道感染	4 (4.0)	4	4	4.0	腹痛、インフルエンザ、悪心及び上気道感染 各4例(4.0%NRBQUOTE(%))
肝機能検査異常	3 (3.0)	3	3	3.0	浮動性めまい及び肝機能検査異常 各3例(3.0%NRBQUOTE(%))

```
/*5.NP_PTList1を転置し、転置後の各変数を「、」で結合して新しい変数JPTList1を作成する*/
/*&Gnp1が0例の場合は、JPTList1に欠損値を格納する*/
```

```
proc transpose data= OUTPUT_D out=OUTPUT_D;
var NP_PTList1;
run;
data OUTPUT_D;
length JPTList1 $32767.;
set OUTPUT_D;
if "&Gnp1"   ="0例" then JPTList1="";
if "&Gnp1" ^="0例" then JPTList1=catx('、',of COL:);
run;
/*6.length check：文字変数のlengthはSAS9.4では32,767バイトが最大のため、CATX関数で連結した記載
（文字変数JPTList1）がそれ以上になる場合、Warningを出力し確認する*/
/*7.記載情報を変数JPTList1に要約しマクロ変数JPTList1に格納*/
data _NULL_;
set OUTPUT_D;
if length(JPTList1) >= 32767 then do;
put "WARNING:JPTList1の文字切れの可能性があります。";
end;
call symputx("JPTList1",JPTList1);
run;
```

⑥ マクロ変数をデータに出力し Compare することで、検証する。新規に定型の文書を作成する場合、入力ストリームに格納する文書テンプレートは MW 部門で確認の取れたテンプレートを格納し、DS/BS 部門でマクロ変数のみをダブルプログラミングにより検証すれば、文書の QC が可能になる。マニュアルで文書作成する際の誤字・脱字や表記ゆれに伴う一貫性チェック等、QC にかける工数を削減できる。

```
/*検証用データセット作成*/
data Main;
G1="&G1"; G2="&G2";   Gnp1="&Gnp1"; Gnp2="&Gnp2";   JPTList1="&JPTList1"; JPTList2="&JPTList2";
run;
```

⑦ 最後に STREAM Procedure を実行することで、定型の文書を半自動で作成できる。

```
/*記載を条件分岐するTemplateマクロ*/
%macro Tmp();
%if "&Gnp1"   ^="0例" and "&Gnp2"   ^="0例" %then %do ;
有害事象は、&G1.&Gnp1.、&G2.&Gnp2.に認められた。 goto NEWLINE;
&G1.で認められた有害事象は、&JPTList1.であった。 goto NEWLINE;
&G2.で認められた有害事象は、&JPTList2.であった。
%end ;
%if "&Gnp1"   ^="0例" and "&Gnp2"   ="0例" %then %do ;
```

```
有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、&G1.のみに認められた。 goto NEWLINE;
&G1.で認められた有害事象は、&JPTList1.であった。
%end ;
%if "&Gnp1"  ="0例" and "&Gnp2" ^="0例" %then %do ;
有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、&G2.のみに認められた。 goto NEWLINE;
&G2.で認められた有害事象は、&JPTList2.であった。
%end ;
%if "&Gnp1"  ="0例" and "&Gnp2" ="0例" %then %do ;
有害事象は、&G1.&Gnp1.、&G2.&Gnp2.であり、認められなかった。
%end ;
%mend Tmp;
/* STREAM Procedureの指定*/
filename out "C:¥temp¥TEAE.txt" lrecl = 32767;
PROC STREAM OUTFILE=out RESETDELIM="goto";
BEGIN goto;
%Tmp()
;;;;
```

4.5. 利用上の注意点

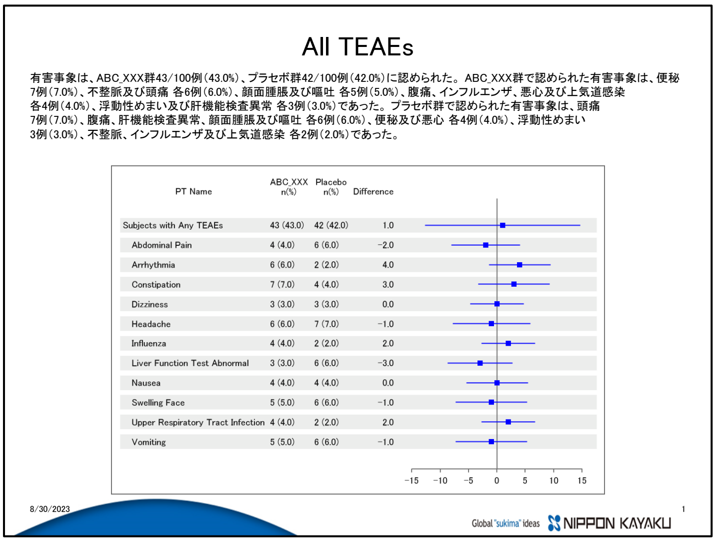
① STREAM Procedure を用いて作成した文書は、テキストファイルとして作成し、書式設定済みの文書テンプレートに Copy&Paste することで活用できる。

② 検討事例は 2 群の PT ごとの集計表に対する定型の文書を作成する仕様とした。もし集計表内に SOC も表示される場合、SOC の indentwt を 0 にすれば、同プログラムで同じ発現 PT のみの記載を作成することは可能である。集計表並びに文書テンプレートの仕様が異なる場合は、カスタマイズが必要になる。

③ 文字変数の length は SAS9.4 では 32,767 バイトが最大のため、超えると Error になる。

④ 日本語の有害事象名に用いる MedDRA/J には「、」を含む基本語もあるため、基本語内の「、」と記載に用いる「、」を区分できない。区分したい場合は、基本語を「」で囲む等の対処が必要である。

⑤ 本稿の目的とは異なるが、吉田（2017）で提案された PowerPoint 形式のプレゼンテーションスライドに、本稿で作成した文書を追加したい場合は、4.4 節の%Tmp から改行コード「goto NEWLINE;」を取り除き、テキスト情報のみのマクロにする。その後 proc odstext 内にマクロとして記載することで、集計表とプロットに加えて、文書を追加することも可能である（右図）。SAS プログラム（Sec4_5.sas）を公開する。



5. その他の検討事例

MW 部門が保有する CSR/CTD/照会事項回答用の文書テンプレートには、DS/BS 部門で作成した集計表の内容について、定型の文書を作成するセクションが、複数存在する。4 章の検討事例の仕様以外についても、一部検討した。簡潔に紹介し、SAS プログラムを公開する。

5.1. その他の検討事例 1

「治験の総括報告書の構成と内容に関するガイドラインについて」（平成 8 年 5 月 1 日付け 厚生省薬務局審査課長通知）セクション 12.2.1 有害事象の簡潔な要約に対する集計表の内容について、CSR の文書テンプレートを参考に、定型の文書を作成する SAS プログラム（Sec5_1.sas）を作成した。

表 12.2.1 有害事象の簡潔な要約／安全性解析対象集団（X 試験）

項目	ABC_XXX N=100 n (%)	Placebo N=100 n (%)
有害事象	43 (43.0)	42 (42.0)
重篤な有害事象	4 (4.0)	1 (1.0)
治験薬との因果関係が否定できない有害事象	7 (7.0)	13 (13.0)

同一症例が、同一の事象を複数回発現した場合、発現者数は1例として算出した。

●は、●/●例（●%）に認められた。

5.2. その他の検討事例 2

「新医薬品の総審査期間短縮に向けた申請に係る CTD のフォーマットについて」（平成 23 年 1 月 17 日付け厚生労働省医薬食品局審査管理課通知）に例示されている有害事象の発現例数（因果関係を問わない全ての有害事象と因果関係が否定できない有害事象で集計した結果を一表に提示）に対する集計表の内容について、CTD の文書テンプレートを参考に、定型の文書を作成する SAS プログラム（Sec5_2.sas）を作成した。

表 X.X 有害事象の発現者数／安全性解析対象集団（X 試験）

	因果関係を問わない		因果関係が否定できない	
基本語	ABC XXX N=100 n (%)	Placebo N=100 n (%)	ABC XXX N=100 n (%)	Placebo N=100 n (%)
有害事象の発現者数	43 (43.0)	42 (42.0)	7 (7.0)	13 (13.0)
腹痛	4 (4.0)	6 (6.0)	2 (2.0)	3 (3.0)
不整脈	6 (6.0)	2 (2.0)	1 (1.0)	1 (1.0)
便秘	7 (7.0)	4 (4.0)	0	1 (1.0)
浮動性めまい	3 (3.0)	3 (3.0)	0	2 (2.0)
頭痛	6 (6.0)	7 (7.0)	1 (1.0)	1 (1.0)
インフルエンザ	4 (4.0)	2 (2.0)	2 (2.0)	0
肝機能検査異常	3 (3.0)	6 (6.0)	0	3 (3.0)
悪心	4 (4.0)	4 (4.0)	1 (1.0)	1 (1.0)
顔面腫脹	5 (5.0)	6 (6.0)	0	0
上気道感染	4 (4.0)	2 (2.0)	0	0
嘔吐	5 (5.0)	6 (6.0)	0	1 (1.0)

MedDRA Version XX.X.

有害事象は、●/●例（●%）に認められた。認められた有害事象は、●● ●例（●%）であった。このうち、●● ●例は、治験薬との因果関係が否定されなかった。

6. まとめ

本稿では、Medical Writer が、有害事象の集計表と文書テンプレートを基に、有害事象名や発現例数、発現割合等の内容についてマニュアルで作成していた定型の文書を、STREAM Procedure を用いて半自動で作成することで Medical Writing を効率化する方法について、解説した。集計表の作成を担う DS/BS 部門が、文書の作成を担う MW 部門の作業プロセスと内容を理解し協力することで、Medical Writing の効率化を推進し、Medical Writer が「人」による高い専門技術や創造性が求められる執筆作業に注力できるようになり、より高品質なドキュメントを短期間に作成することが可能になる。以上より、集計表に基づく定型文書の効率的な作成手段として、STREAM Procedure を用いた方法は今後、有用な Medical Writing の支援ツールになり得ると期待される。

7. 参考文献

[1] Joseph H. Proc STREAM: The Perfect Tool For Creating Patient Narratives. Proceedings of the Pharmaceutical SAS Users Group (PharmaSUG) Conference. inVentiv Health, Princeton, NJ, 2015. Available at https://www.pharmasug.org/proceedings/2015/AD/PharmaSUG-2015-AD03.pdf.

[2] Joseph H. The New STREAM Procedure as a Virtual Medical Writer. Proceedings of the Pharmaceutical SAS Users Group (PharmaSUG) Conference. inVentiv Health, Princeton, NJ, 2016. Available at https://www.pharmasug.org/proceedings/2016/AD/PharmaSUG-2016-AD17.pdf.

[3] Joseph H. Making Documents 'Intelligent' with Embedded Macro Calls, DOSUBL and Proc STREAM: An example with the CONSORT Flow Diagram. Proceedings of the Pharmaceutical SAS Users Group (PharmaSUG) Conference. inVentiv Health, Princeton, NJ, 2017. Available at https://www.lexjansen.com/pharmasug/2017/BB/PharmaSUG-2017-BB02.pdf.

[4] SAS Institute Inc. "STREAM Procedure." Base SAS(R) 9.4 Procedures Guide. Date accessed: 13July2023. SAS Institute Inc., Cary, NC, Available at https://documentation.sas.com/doc/en/pgmsascdc/v_041/proc/n1sak3n3asxfbqn1aw24lxdvez69.htm

[5] SAS Institute Inc. "STREAM Procedure." Base SAS(R) 9.4 Procedures Guide. Date accessed: 13July2023. SAS Institute Inc., Cary, NC, Available at https://documentation.sas.com/doc/ja/sasstudiocdc/v_041/pgmsascdc/proc/n12zrkr08eiacmn17lcv4fmt79tb.htm

[6] 大橋靖雄. 日本メディカルライター協会 第 18 回シンポジウム要旨 2019, http://www.jmca-npo.org/seminar/20191128.html（2023 年 8 月 31 日アクセス可能）

[7] 厚生省薬務局審査課長通知. 薬審第 335 号 治験の総括報告書の構成と内容に関するガイドラインについて 1996, https://www.pmda.go.jp/files/000156923.pdf（2023 年 8 月 31 日アクセス可能）

[8] 厚生労働省医薬食品局審査管理課通知. 事務連絡 新医薬品の総審査期間短縮に向けた申請に係る CTD のフォーマットについて 2011, https://www.pmda.go.jp/files/000209183.pdf （2023 年 8 月 31 日アクセス可能）

[9] 関根暁史. proc STREAM による Analysis Results Metadata の作成. SAS ユーザー総会 論文集 2017, 407-422.

[10] 山崎文寛. PROC STREAM を用いた Analysis Data Reviewer's Guide の効率的な作成方法の提案. SAS ユーザー総会 論文集 2018, 133-153.

[11] 吉田直記・山崎文寛・舟尾暢男・高浪洋平. SAS による臨床試験の解析速報に用いるプレゼンテーションスライドの効率的な作成方法の提案. SAS ユーザー総会 論文集 2017, 339-360.

8. 連絡先

E-mail : takayuki.hirai@nipponkayaku.co.jp

SASにおけるデータ処理の基礎 複数データの結合と構造転換

○山野辺浩己

（マルホ株式会社臨床開発部）

データの結合や構造転置といったデータハンドリングは、解析を行うにあたり必要な前処理であり、また出力のための後処理で、必須の工程である。SASにおけるデータハンドリングには、ソースとなるデータセットを1行ずつ読み込みPDV(プログラムデータベクトル)上に展開して処理しその結果を出力するという特徴があり、その性質を生かしてデータ構造の展開が可能である。

SASにおけるデータ処理の特徴は、上述の通りPDV上で行われる行単位の処理である。この処理の特性を活かして、Mergeステートメントを使い、複数のデータセットを横方向に結合し、１つのデータセットを作成することを可能にしている。しかし、複数のデータセットのデータをPDVに流し込み処理するにあたり、ユーザーが想定しない処理が引き起こされ、誤った解析結果を導かれることもある。複数データセットを縦結合するに用いられるsetステートメントにおいても同様に、PDV上で処理する都合、ユーザーの想定とは違った挙動を引き起こされることがある。

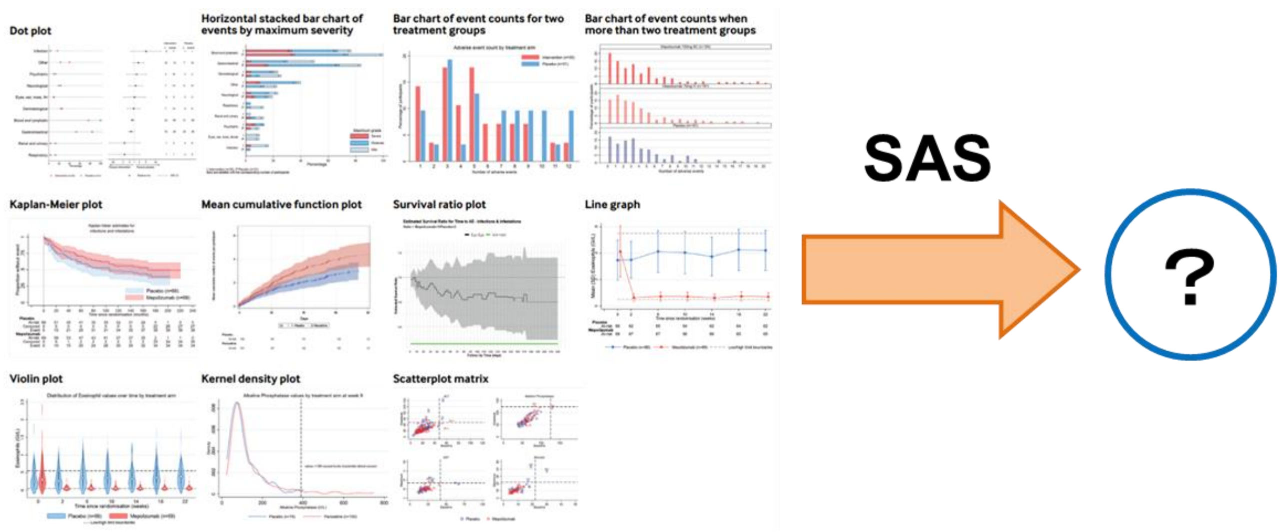
本発表では、ビギナーを対象に、データの結合や構造転置といったデータハンドリングの基礎部分を説明し、実用的なコードと陥りがちな失敗事例について紹介する。

ランダム化比較試験における有害事象のSASによる視覚化

○小山田隼佑[1]、徳田芳稀[2]

（[1]NPO法人JORTC、[2]エイツーヘルスケア株式会社）

CONSORTを含む様々なランダム化比較試験(RCT)の結果報告ガイドラインでは、RCTにおいて出現した有害事象の分析の際、視覚化によって有害プロファイルを要約して報告することが奨励されている。しかし、実際に有害事象の視覚的要約を利用している雑誌論文は僅かである(Phillips R et al, BMJ Open 2019)。そのような背景から、Phillipsらは主に医薬品の治験に焦点を当て、治療群間の安全性の比較に際し推奨された有害事象のデータを視覚化するための10のプロットに対する解釈と推奨事項、および視覚化の方法の選択に有用な決定木に関する論文を発表している(Phillips R et al, BMJ 2022)。本論文内において、各プロットを統計ソフトでどのように再現するかについて言及があるが、実際のSASコードは公表されていない。そこで本発表では、推奨されている10のプロットについて説明するとともに、SASでどこまで再現できるのかを報告する。



Phillips R et al, BMJ 2022

【参考文献】

Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open, 2019;9:e024537.

Phillips R, Cro S, Wheeler G, Bond S, Morris TP, Siobhan Creanor S, et al. Visualising harms in publications of randomised controlled trials: consensus and recommendations. BMJ, 2022;377:e068983.

前処理大全

SASバージョン

森岡　裕

（イーピーエス株式会社）

Awesome SAS Data Handling Techniques

Yutaka Morioka

(EPS Corporation)

書籍　前処理大全［データ分析のためのSQL/R/Python実践テクニック］(2018年4月13日出版，本橋智光　著，株式会社ホクソエム　監修，技術評論社)　はデータの前処理に焦点をあて，SQL/R/Pythonでの実践的なコードを並列で掲載するとともに，様々な観点から総合的により良いコード（Awesomeコード）とそうではないコードの解説を行っている．データ解析を生業とする万人にとっての名著であることは疑いがないが，一つ残念な点がある．SASコードが入っていないことである．そこで，前処理大全の目次トピックからチョイスして，SASで前処理を書くならどうするか，Awesomeな書き方とは何かについて私見を述べたい．

とりあげるトピックは以下となる．

第2章 抽出
2-1 データ列指定による抽出
2-2 条件指定による抽出
2-3 データ値に基づかないサンプリング
2-4 サンプリング

第3章 集約
3-1 データ数，種類数の算出
3-2 合計値の算出
3-3 極値，代表値の算出
3-4 ばらつき具合の算出
3-5 最頻値の算出
3-6 順位の算出

第4章 結合
4-1 マスタテーブルの結合

第7章 展開
7-1 横持ちへの変換

第8章 数値型
8-1 数値型への変換
8-4 正規化
8-7 数値の補完

第9章 カテゴリ型
9-1 カテゴリ型への変換
9-2 ダミー変数化

第12章 位置情報型
12-2 2点間の距離，方角の計算

参考文献

本橋智光　著，株式会社ホクソエム　監修，技術評論社，前処理大全［データ分析のための SQL/R/Python 実践テクニック］(2018 年 4 月 13 日出版)

小さく始めるSGPLOT／SGPANEL

～データに語らせよう～

○太田　裕二、浜田　泉、石川　優子、森田　祐介、南雲　幸寛、西部　莉央

（ノーベルファーマ株式会社　データサイエンス部）

データの可視化に対するニーズや関心が高まっている。データを集める時代からデータが集まる時代へ、身近にデータがあふれる時代となった。このため、データから価値ある情報を導く解析のニーズも一段と増している。また、データの可視化による品質管理（データクリーニング）も注目されている。

SGPLOTプロシジャとSGPANELプロシジャは、ODS Graphics機能を活用してグラフを作成するプロシジャである。箱ひげ図、散布図、スパゲティプロット、折れ線グラフ、棒グラフ、ヒストグラムなど様々な種類のグラフを柔軟に作成することが可能で、過去のSASユーザー総会でも多数の研究発表を見つけることができる。

SGPLOT／SGPANELプロシジャはデータの可視化に有用なプロシジャである一方で、その多機能さ、柔軟さ故に、初学者にはハードルが高い面がある。

そこで本発表では、見栄えは深く追求せず、初学者がSGPLOT／SGPANELプロシジャを使ったグラフ作成を小さく始めるために必要なポイントに限定して紹介する。最初にグラフを任意の外部ファイルとして出力する方法を、次にSGPLOTプロシジャで軸、シンボル・線及び凡例を調整する方法と実務でよく使用されるいくつかの種類のグラフを作成する方法を紹介する。その後、SGPANELプロシジャでグラフを作成する方法と併せてSGPANELプロシジャの使用が推奨される状況も紹介する。

第1段階と第2段階で異なる2値評価項目を用いた アダプティブシームレスデザインに対する仮説検定法の実装

○高橋　健一[1]、石井　亮太[2]、丸尾　和司[2]、五所　正彦[2]

([1]MSD株式会社、[2]筑波大学)

アダプティブシームレスデザインは、複数の治療群と対照群を比較し最適な治療群を選択するための第II相試験（第1段階）と、第II相試験で選択された治療群と対照群を比較する第III相試験（第2段階）を組み合わせた臨床試験デザインである。2つの試験を1つの試験で実施するため、試験期間の短縮や第1段階と第2段階のデータを併合することでより少ない被験者で試験を実施できる利点があるため、研究者や製薬企業にとって魅力的なデザインである。本発表では、以下のアダプティブシームレスデザインを対象とする。第1段階終了後の中間解析では、主要評価項目よりも短期間で測定可能な2値の評価項目を用いて、最も有効な治療群を選択する。第2段階終了後の最終解析では、2値の主要評価項目を用いて、選択された治療群と対照群を比較する。ここで、中間解析での選択を考慮せずに最終解析で検定を実施すると、第一種の過誤確率の増大を招く。Takahashi et al. (2022)では、最終解析の検定として「中間解析で最も効果の高い治療群が選ばれた」という条件の下での条件付き分布を用いた正確検定及びmid-p値に基づく検定を提案した。本発表では、Takahashi et al. (2022)での提案法の実装方法を紹介する。

参考文献

- Biswas A, Hwang JS. A new bivariate binomial distribution. Statistics & Probability Letters. 2002; 60:231-40.
- Takahashi K, Ishii R, Maruo K, Gosho M. Statistical tests for two-stage adaptive seamless design using short- and long-term binary outcomes. Statistics in Medicine. 2022; 41:4130-42.

臨床統計解析ならびにRWD活用のための新しいSASソリューション：SAS® Health Clinical Acceleration / Cohort Builderのご紹介

○土生 敏明[1]、William Kuan[1]

（[1]SAS Institute Japan）

昨今，Real World Data の活用や DCT・eCOA といったデータを用いて治験の効率化・高度化が求められてきている. SAS は過去より現在まで医薬の臨床領域において様々なソリューションを提供しており，DCT や eCOA を含めた臨床データ管理・活用には SAS Health Clinical Acceleration を，RWD においては SAS Health Cohort Builder といったソリューションを提供している.

SAS Health Clinical Acceleration は Cloud にて提供するソリューションだが, 様々な規制要件を満たせれるデータレポジトリとして，バージョンコントロール・版管理機能・電子署名機能・Audit Trail 機能を先駆けてリリースし、R/Python といったプログラミング機能を今後拡張していく予定としている.

Cohort Builder は, Real World Data を誰でも活用できるようにする為のソリューションであり, 様々な Real World Data を SAS 標準形式にマッピングし，GUI にてドラッグアンドドロップやプルダウンにてデータ選択，選択基準/除外基準等を設定できる．また解析テンプレートを作成・登録することにより，同じ解析を別の方が別のデータを用いて実施する事が出来る．これら結果は Visual 化/AI による自動分析等も可能である.

これら製品が臨床試験・臨床研究に携わる方々の業務に対する一助を担いたく製品概要説明をする.

CLASSDATAオプションを利用した基本的な集計方法について

森岡　裕

（イーピーエス株式会社）

Happy Summary Life with Classdata

Yutaka Morioka

(EPS Corporation)

Proc Summary/Meansにおいて，Classdataオプションを利用することで，実際にデータとして生じていない集計水準を０件として出力することができる．これによって，データの出力レイアウトを，データセットの状態によって，都度調整する必要がなくなり，集計が便利になる．

他にも集計軸の設定に関連したオプション(nway, types, ways, missing, preloadfmt, completetypes, exclusive)について紹介し，基本的な要約統計・頻度集計のプログラムについて，今一度見つめなおしたい．

```
data clds;
do AGE=10 to 16;
 do SEX="男子","女子";
       output;
 end;
end;
run;
```

VIEWTABLE: Work.Clds

	AGE	SEX
1	10	男子
2	10	女子
3	11	男子
4	11	女子
5	12	男子
6	12	女子
7	13	男子
8	13	女子
9	14	男子
10	14	女子
11	15	男子
12	15	女子
13	16	男子
14	16	女子

先にフルセットの水準をデータセットで作成したのち，それをclassdata=オプションで指定することで，その水準で集計が行われる

```
proc summary data=sashelp.class classdata=clds nway ;
   class AGE SEX;
   var weight;
   output out=out1 n= mean= std= min= median= max=/
  autoname;
run;
```

VIEWTABLE: Work.Out1

	Age	Sex	_TYPE_	_FREQ_	Weight_N	Weight_Mean	Weight_StdDev	Weight_Min	Weight_Median	Weight_Max
1	10	女子	3	0	0	.	.	.	.	.
2	10	男子	3	0	0	.	.	.	.	.
3	11	女子	3	1	1	50.5	.	50.5	50.5	50.5
4	11	男子	3	1	1	85	.	85	85	85
5	12	女子	3	2	2	80.75	5.3033008589	77	80.75	84.5
6	12	男子	3	3	3	103.5	22.765104876	83	99.5	128
7	13	女子	3	2	2	91	9.8994949366	84	91	98
8	13	男子	3	1	1	84	.	84	84	84
9	14	女子	3	2	2	96.25	8.8388347648	90	96.25	102.5
10	14	男子	3	2	2	107.5	7.0710678119	102.5	107.5	112.5
11	15	女子	3	2	2	112.25	0.3535533906	112	112.25	112.5
12	15	男子	3	2	2	122.5	14.849242405	112	122.5	133
13	16	女子	3	0	0	.	.	.	.	.
14	16	男子	3	1	1	150	.	150	150	150

参考文献

SAS 忘備録「PROC MEANS の、COMPLETETYPES・PRELOADFMT オプションの紹介」

https://sas-boubi.blogspot.com/2015/03/proc-meanscompletetypespreloadfmt.html

(Accessed Aug 18,2023)

SASによる散布図行列の実装

○徳田　芳稀

（エイツーヘルスケア株式会社）

散布図行列とは、複数の変数を持つデータについて、主に散布図やその確率楕円等を行列の形式で表示することにより、複数の変数間の関連を一度に可視化できる図である。SASでは、SGSCATTERプロシジャによりデータセット内の2変数間の散布図やその確率楕円等、更に各変数のヒストグラムやカーネル密度等を含めた散布図行列の作成が可能である。一方OSS(Open Source Software)の1つであるRでは、GGallyパッケージ等を用いて散布図行列の作成が可能である。SASのSGSCATTERプロシジャ及びRのGGallyパッケージそれぞれで作成した散布図行列を比較すると、Rによる散布図行列の方が散布図やカーネル密度だけではなく、相関係数等の多様な情報を描画可能である。SASのSGSCATTERプロシジャを用いて、Rと同様の散布図行列を作成するには、オプション等の工夫だけでは限界があり、更に描画不可能なものも存在する。そこでSAS GTL(Graph Template Language)のlayoutステートメント及びplotステートメントを利用し、Rによる出力と同等の描画内容を、SASにより実装する方法を紹介する。

アダプティブ臨床試験の動作特性を測るシミュレーションの実践

○中村 将俊[1,8]，青木 誠[2,8]，飯塚 政人[3,8]，高津 正寛[4,8]，田中 勇輔[5,8]，棚瀬 貴紀[6,8]，菅波 秀規[7,8]

([1]ファイザーR&D合同会社，[2]ノバルティスファーマ株式会社，[3]田辺三菱製薬株式会社，[4]持田製薬株式会社，[5]アステラス製薬株式会社，[6]大鵬薬品工業株式会社，[7]興和株式会社，
[8]日本製薬工業協会 医薬品評価委員会 データサイエンス部会）

本文

ICH-E20のトピックとして「アダプティブ臨床試験」が採択され，ガイドラインの作成が検討されている．2023年7月時点において，European Medicines AgencyとFood and Drug Administration（FDA）からアダプティブデザインのガイダンスが発行されている．FDAガイダンスでは，アダプティブデザインを「臨床試験に参加した被験者の蓄積されたデータに基づいて，試験デザインの1つ以上の側面について，予め計画された変更を行うことができる臨床試験デザイン」と定義している．アダプティブデザインにより与えられる柔軟性により，試験の参加者がより良い治療を受けられる機会が増え，より効率的な医薬品開発，さらには試験リソースの節約といった利点が期待される．その一方，統計手法を適切に用いなければ，第一種の過誤確率の増大，点推定値へのバイアスの発生，信頼区間の被覆確率が名義水準と異なるなど，統計的妥当性の観点から望ましくない現象が起こる可能性がある．日本製薬工業協会医薬品評価委員会データサイエンス部会は，アダプティブデザインの適切な実施を促進するために，FDAから公表されているアダプティブデザインに関するガイダンスの邦訳と，アダプティブデザインの基本的な統計的推測法に関してまとめた「アダプティブデザインの統計的推測に関する検討」を公表している．

FDAはさらに近年，「Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products」というガイダンスを公表し，革新的な試験デザイン（CID）の取り組みを進めている．FDAは，CIDに固定の定義は存在しないが，この試験デザインの事例の一つとしてアダプティブデザインを挙げている．また，「多くのCIDに共通する一つの特徴は，試験の動作特性を推測するためには数式よりもシミュレーションが必要となることである 」と述べていることから，アダプティブデザインの動作特性を理解するためにはシミュレーションが重要であることは規制当局も認識していると考えられる．

新しい臨床試験を計画するには必要となるいくつかのステップがある．本発表は，アダプティブデザインを立案する場合に求められる動作特性を評価するためのシミュレーションについて，その方法と手順，シミュレーションで評価すべき指標とその結果のまとめ方について，主にMayer et al. (Stat Biopharm Res 2019, 11, 4, 325–335) の論文及びFDAガイダンスを参考にし，まとめた内容を報告する．

SASで始めようconstrained Longitudinal Data Analysis

～君たちはベースライン値をどう扱うのか～

〇森田　祐介、太田　裕二、浜田　泉

（ノーベルファーマ株式会社　データサイエンス部）

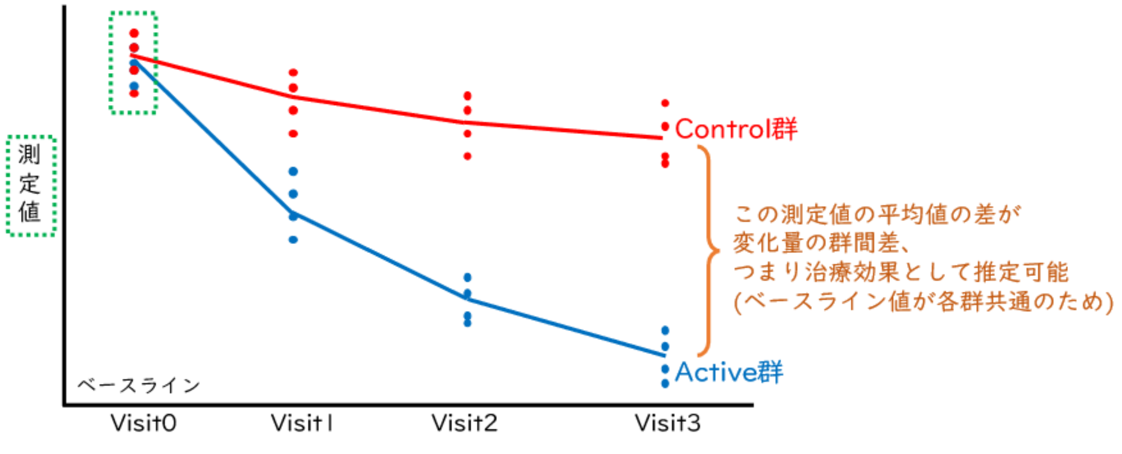
臨床試験では、被験者の経時的な反応プロファイルを把握するため、反復測定がしばしば行われる。これらの経時的なデータは、通常、被験者内相関を持つとされ、その相関を適切に取り扱うことが解析の要点である。さらに、欠測値への対処も重要となる。このため、反復測定データ解析では、ベースライン値からの変化量を応答変数として、時点、群、そして時点と群の交互作用項を固定効果、ベースライン値を共変量とし、誤差項の被験者内相関に無構造を仮定するMixed Model for Repeated Measures(MMRM)解析が多く採用されている。一方で、ベースライン値も応答変数として扱い、無作為化を前提に、介入前のベースライン値を各群で共通とする制約(constrained)の下で解析するconstrained Longitudinal Data Analysis(cLDA)も提案されており、日本の医療用医薬品の審査報告書に、その実例も確認できる。

本発表では、MMRMとの比較を含めて、cLDAの特性について紹介するとともに、SASにおける実装例を示す。特に、ベースライン値に欠測が頻発する状況では、cLDAがMMRMに比べて優れた検出力を持つとの報告もあり、生物統計担当者は、状況に応じたベースライン値の取り扱いを考慮する必要がある。

cLDAとは？

Nobelpharma ノーベルファーマ株式会社

- Liang and Zegerが2000年に提案
- 応答変数は変化量ではなく測定値とする
- ベースライン値は共変量ではなく応答変数とする
- 無作為化しているため、ベースラインの平均値は各群共通とする(constrained)
- ベースラインより後の各時点では投与群ごとに平均値が異なると仮定する



SASのSGPLOTプロシジャを用いたデータ可視化入門

○五味隆佑

（コムチュア株式会社）

SGPLOTプロシジャは、データの可視化を行う際の幅広いサポートを提供している。このプロシジャでは、基本的なグラフから応用的なものまで、多岐にわたるグラフの作成が可能である。

本資料では、SGPLOTプロシジャの詳細な使用方法と機能に焦点を当てて解説する。SGPLOTの全体的な概要とその主要な特徴を始めに紹介する。さらに、グラフを描画するためのステートメントや、近似式（LOESS近似や多項式近似）や参照線を追加するためのステートメント、そしてグラフの細部を調整するためのオプションやステートメントについて解説する。

本資料は、SGPLOTプロシジャを使用したデータ可視化の手引きとして構築されている。前半部分では、SGPLOTの基本的な概要とその中で利用可能なステートメントを、グラフを描画するステートメント、近似式や参照線を追加するステートメント、そしてグラフの軸や凡例を調整するステートメントの三つに分けて解説する。その後、様々なグラフとそれらを生成するためのコードを具体的に示す。

後半では、本題のデータ可視化入門として、特に使用頻度の高い棒グラフ、折れ線グラフ、散布図、ヒートマップ、ヒストグラム、箱ひげ図を描画方法に焦点を当てる。それぞれのグラフが最も適しているデータの種類やより洗練されたグラフを作成するためのオプションや調整方法も詳しく説明する。最後に、グラフの軸調整に関するステートメントについての補足説明を加える。

SASアカデミックプログラムご紹介

絹谷　明

（SAS Institute Japan株式会社　エデュケーション部　部長）

本プレゼンテーションの目的は、SASが大学・教育機関の教職員ならびに学生向けに提供するサービスを紹介し、活用を働きかけることである。そのサービスには、教育機関とSASが共同で単位認定するプログラムの種類と特典、教職員および学生向けの教育・学習リソース、無償のSASソフトウェア、およびデジタルバッジが含まれる。

冒頭では共同単位認定プログラムであるSAS Academic Specializationについて触れ、その概要およびTier1からTier3までの3段階それぞれに求められる要件と特典を解説することで、同プログラムへの関心の喚起を促す。

Academic Specializationプログラムに続き、一般的に教職員ならびに学生向けに提供されるSASの無償またはアカデミック割引のサービスについても紹介する。教員向けに提供されるSAS Educator Portalというポータルサイトは教育に使える資料やデータセット、無償で利用できるe-Learningなどの教育リソースや割引制度を含み、学生向けに提供されるSAS Skill Builder for Studentsは同様のe-Learningや学生向けの資格割引案内を含む。

続いてSASのグローバル認定制度を紹介し、日本語で受験可能な試験を示すとともに、グローバル資格を取得することの意義およびメリットを解説する。試験料はアカデミック向けの割引制度も活用可能なため、より多くの教育機関での受験を推奨する。

次にSASが教職員・学生向けに無償で提供するソフトウェアであるSAS Viya for LearnersおよびSAS OnDemand for Analyticsを紹介し、それぞれの用途ならびに特徴を解説する。教員の指導用や自習用途、また学生による自習教材や研究用のために活用を広く働きかける。

最後にAcademic Specializationやグローバル認定資格、e-Learningなどで取得できるデジタルバッジに関する概要および使用方法について解説し、デジタルバッジの活用による取得者のメリットを紹介する。

プレゼンテーションの発表としては以上となり、資料の参照用に発表中に紹介した無償ポータルやソフトウェアなどの登録方法の手順を、教職員向けおよび学生向けそれぞれに用意した。

以上

SASにおける外部APIと自然言語の利用例

○中松 建

（個人）

背景：2023年は第4次AIブーム始まりの年とも言われ、そのうち大規模言語モデル(LLM)については各種プログラミングへ大きな影響があるとの予想も多い。各種システムでLLMを利用する方法のひとつにはAPIが挙げられ、またDX・RPA関連やHL7 FHIRなど外部APIを使用するような機会も多くなっている。
そこでBase SASの機能からOpenAI社のAPI実行を確認し、続けてLLMとSASの連携を検討した。

確認・検証：OpenAI社のAPIを対象としてHTTPプロシージャ・LIBNAME JSONエンジンおよびDS2プロシージャ(http, jsonパッケージ)によるAPIの実行を確認した。外部APIからLLMにより自然言語が扱えるようになったことから、次に自然言語によるSASプログラムの実行を検討し、OpenAI社APIの関数呼び出し機能およびSASで作成したlangchainを模したエージェントにより、自然言語による入力内容に応じたSASマクロの実行を確認した。

まとめ・結果：SASから外部APIの利用は、入出力用のJSONファイルの扱いについて少し習熟や工夫も必要ではあるものの、SASマクロやSTREAMプロシージャなどを活用すれば、エージェントのような動的なプロンプトの生成・処理も可能であった。自然言語を入力としたプログラムの実行ではデータ処理や条件などの可読性も向上することから、自然言語の利用はSASプログラミングにおいても新しい方向性のひとつとなることが考えられる。

GitHubリポジトリ：https://github.com/k-nkmt/SAS_API_LLM_Examples

営業活動効果分析ツールの開発事例紹介

○佐藤耕一

（株式会社タクミインフォメーションテクノロジー）

営業活動における活動内容の効果検証、営業活動の課題把握と課題解決のためのアクションを効率的に分析するための営業活動効果分析ツールの開発事例を紹介します。本ツールは営業活動内容のバリエーションごとに説明変数を作成し、営業成果を目的変数としたロジスティック回帰モデルを構築します。また.NET VBで構築したユーザーインターフェース画面をSAS Enterprise GuideにAdd-in登録して使用し、説明変数の作成、データ加工、データ分析のエンジンにSASを利用したシステム構成とすることにより、ロジスティック回帰モデルによる営業活動内容の効果検証や構築したモデルを利用したシミュレーションを効率よくノンプログラミングで実施できます。さらにSAS Enterprise Guideのタスクを並行して利用することができる枠組みとしたことにより、分析過程に発生する補助的なデータ加工やデータ分析もノンプラグミングにて実施することができます

疫学研究でよく用いるSASプロシジャの紹介

○矢田 徹

（イーピーエス株式会社）

要旨：

疫学研究でよく用いる指標を計算するためのプロシジャを紹介し、いろんな条件での算出結果を示してみる。

キーワード：

Epidemiology, Odds Ratio, Risk Ratio, Risk Difference, Standardized Difference

オッズ比、リスク比、リスク差、率比、標準化差は疫学研究でよく用いられる指標であり、SASにはこれらを効率的に計算する様々なプロシジャがある。単純なモデルでのオッズ比はいくつかのプロシジャのいずれを用いても算出できるが、多変量モデルや重みを考慮する場合などでは、適切なプロシジャやオプションの設定が必要になる。共変量や重みの有無、分布等の条件をいろいろ変えて計算するプログラム例を紹介する。LOGISTICプロシジャ、STDRATEプロシジャ、GENMODプロシジャ、CAUSALTRTプロシジャ、PSMATCHプロシジャを用いる。

SAS Viyaによるリアルワールドデータの効率的利活用

○平井 岳大、古藤 諒、堀江 義治

（アストラゼネカ株式会社　データサイエンス部）

背景　製薬企業では承認申請、製造販売後データベース（DB）調査、Evidence generation等、様々な目的でリアルワールドデータ（RWD）を利用する機会が増えておりその重要性が増している。しかし、RWDは大規模で、治験を実施した際に入手できるデータのように標準化されておらず、データの取り扱いに十分な検討と定義づけが必要である。解析目的に応じて、より速くかつ質の高い解析結果の作成とその結果の解釈、および検討に必要な時間を確保するため、どのように品質を管理するかということも十分な検討が必要である。

目的　大規模なデータを効率よく解析するためにSAS Viyaを利用してTableとFigure（TF）の標準化による品質の向上と効率化について検討する。

方法　次の①から⑤で示す手順で進めた。

① 解析環境の最適化（SPREとCASの効率的利用）

② 過去のDB研究で使用されたTFの調査

③ 標準化が可能なTFの選択

④ 標準TFカタログとそれを作成するための標準SASプログラムの作成

⑤ DB研究で標準TFカタログを利用と効果の定量化

解析ソフトウェアとしてSAS Viya 3.5を使用する。

結果　RWDの大規模なデータ解析を実行する際、SAS ViyaでCASを利用することによってSPREを使うより、データロード処理は約4.5時間が約2.5時間になり、データ加工処理は約5.5時間が約2.5時間になり、処理速度が大幅に改善することを確認した。次に、過去のDB研究で使用されたTFとデータサイエンス部内での議論から標準化が可能な17 TFを選んだ。標準TFカタログを作成し、それらを作成するための標準SASプログラムを作成した。内訳として、Patient disposition、Patient background、発生割合、発生率、サブグループ解析、カプランマイヤー曲線、Cox回帰、ロジスティック回帰、線形回帰、累積発生率曲線、Fine-Gray回帰に関するTFが含まれる。

結語　RWDを解析するツールとしてSAS Viya上でCASを利用することは時間を効率的に利用する上で有効であり、TFの標準化によって質の高い解析結果を作成することができる可能性を有している。これにより、解析計画の検討や解析結果の解釈、および検討するために十分な時間をとることが可能となる。

主催：SAS ユーザー会　世話人会

代表世話人	**伊藤 陽一**	**北海道大学**
世話人(50 音順)	上村 鋼平	東京大学
	岸本 淳司	九州大学 ARO 次世代医療センター
	菅波 秀規	興和株式会社
	林 行和	エイツーヘルスケア株式会社
	森岡 裕	イーピーエス株式会社

SAS ユーザー総会 2023 についての問い合わせ先：info@sas-user2023.ywstat.jp

協賛(50 音順)

イーピーエス株式会社
エイツーヘルスケア株式会社
ClinChoice 株式会社
SAS Institute Japan 株式会社
SAS Institute Japan 株式会社 JMPジャパン事業部
株式会社新日本科学 PPD
スタットコム株式会社
株式会社タクミインフォメーションテクノロジー
株式会社日立製作所

SAS ユーザー総会 2023 論文集ー初版　作成日 2023-09-08