動画による統計表現

～SASはRのグラフをまとめて動画にする～

関根　暁史
エイツーヘルスケア株式会社／生物統計部

It's easy to create dynamic graphs in SAS using the R graphics.

Satoshi Sekine
A2 Healthcare Corp./Biostatistics Dept.

要旨

SAS は R で作成した静止画をまとめて動画にしてしまう。これにより R ならではの美しい画質の動画を作成することが可能となった。

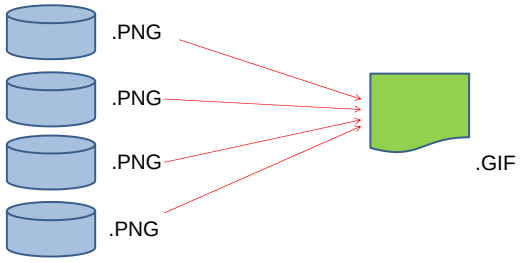
キーワード：R，SAS グラフ，PNG，GIF，ANNOTATE

1. はじめに

昨年度の発表にて SAS は複数の静止画(PNG ファイル)をまとめて動画(GIF ファイル)にすることを示した [1)]。PNG ファイルは SAS 由来である必要もなく、今年度は R 由来のグラフィックをまとめて動画とすることを考案する。具体的には SAS が xpt ファイルの形で R にデータを引き渡し、R の関数が PNG ファイルを生成し、その PNG ファイルを SAS が回収して動画とする。この一連の動作は SAS が制御するものとする。これにより R の美しい画像を動画とすることができるし、R 独特の関数を SAS 側が使うことも可能となる。

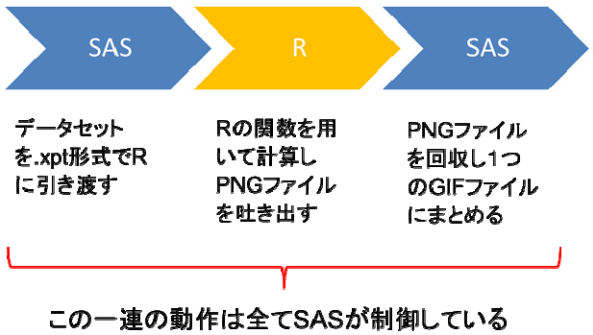
本論文中の動画は、当日発表用スライド（パワーポイント）をご参照頂きたい。

2. 今回のコンセプト



SAS の動画作成のイメージ

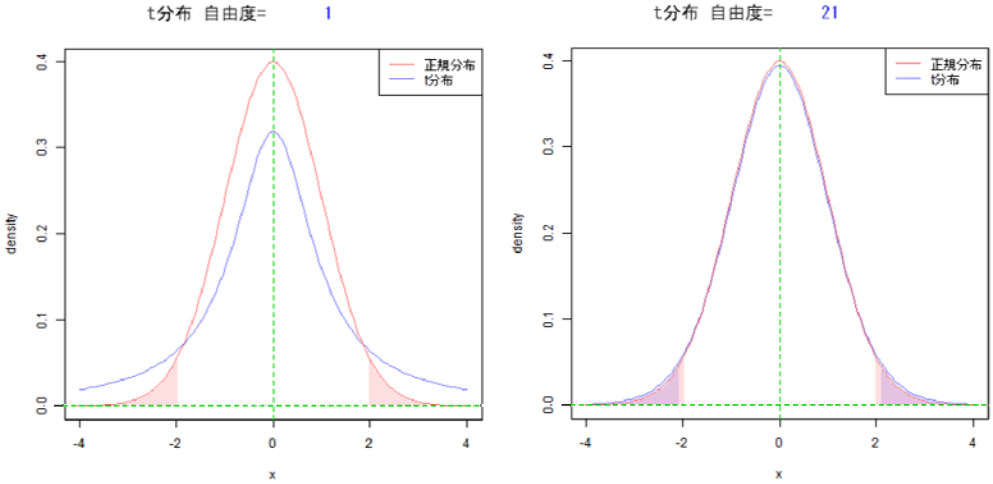
SAS の動画作成システムはコンベンショナルな G 系グラフ(GPLOT，GANNO 等)の機能である [2)]。一方 PNG ファイル(静止画)は ANNOTATE データセット化することが可能なので、ANNOTATE データセットを GANNO プロシジャなどで回転させてしまえば GIF ファイル(動画)は作成可能となる。



今回のフロー

つまり PNG ファイルは G 系グラフ由来のものである必要もなく、SG 系グラフ由来でも R 由来でもよいということになる。今回はこの ANNOTATE 機能を利用して R の美しいグラフを動画にするものとする。今回のフローを左図に示したが、xpt ファイルの形式で R にデータを引き渡し、R の関数が PNG ファイルを生成し、その PNG ファイルを SAS が回収して動画とする。この一連の動作は全て SAS が制御しているが、SAS で R を操作する方法は参考文献 3)を参照頂きたい。上記の方法により、R でのデータハンドリングが不得手である人でも R に引き渡す直前まで SAS でハンドリングすることができるし、SAS にはない R の関数を使用することも可能である。以下の論文の用語であるが、プロシジャと表現した場合は SAS が動作していて、パッケージ・関数と表現した場合は R が動作しているものとお考え頂きたい。

3. t 分布と自由度



（動画の始まり）　　（動画の終わり）

図 1. t 分布と自由度

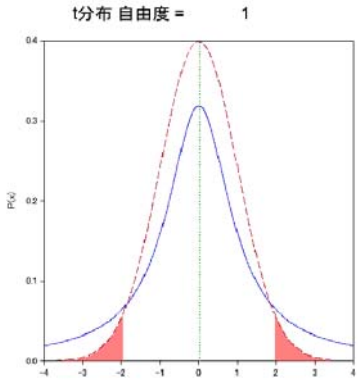


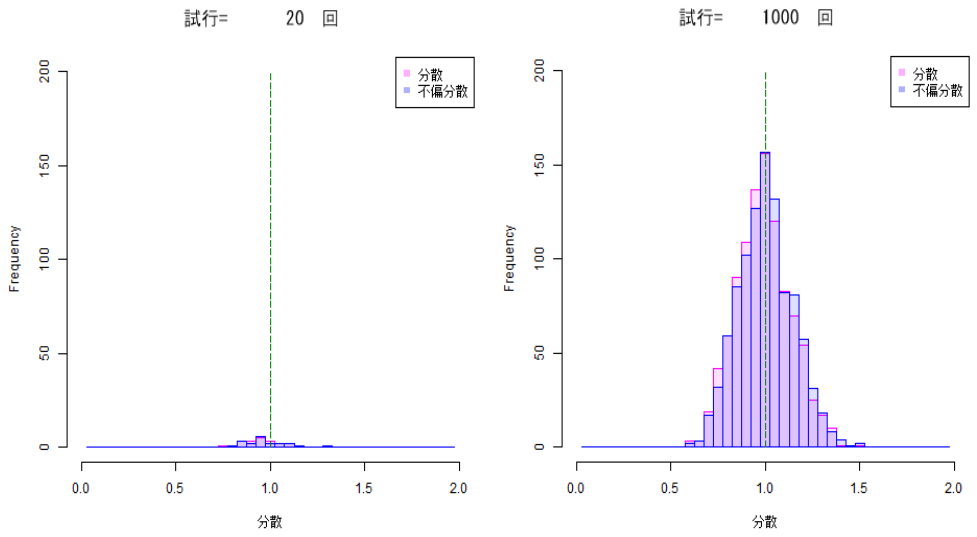
図 1-2. t 分布と自由度(SGPLOT 使用)

t 分布の自由度を増大させると正規分布に漸近していく動画を作成した(図 1)。R は異なる複数のグラフを重ね描きするのが得意である。ヘッダーの「t 分布　自由度=○」の部分は SAS 側から描き入れている。

ここで、R を用いなくとも SGPLOT を用いればそれなりに綺麗な絵ができるのではないかと思われる方がいるかも知れない。ここで R の機能を用いずに SGPLOT のグラフで同様の動画を作成してみたが(図 1-2)、動画作成の過程でグラフの縦横比が微妙に崩れてしまい、どうしてもぼやけた絵になってしまう。今回 R を用いているのは、その理由が大きい。

4. 分散と不偏分散

分散が不偏分散よりも推定値を小さく見積もっていることを視覚的に表現することを考えた。先ず正規分布(平均 0, 標準偏差 1)に従う母集団を作成し、その母集団から surveyselect プロシジャにて 100 レコードずつ無作為非復元抽出を繰り返し 1000 組のセットを作る。その 1 組(100 レコード)の分散と不偏分散を計算してはヒストグラムとして積み上げていく動画を作成した(図 2)。R では色を半透明にできるので、重なり合ったヒストグラムの形状を比較することができる。これによって分散が不偏分散よりもやや低く見積もっていることが分かる。但し正規乱数や surveyselect のシードの値やヒストグラムの分割の仕方によっても見え方は変わってくるので、"不偏分散の方が分散よりも正しい"とはこのグラフのみでは言い難い。



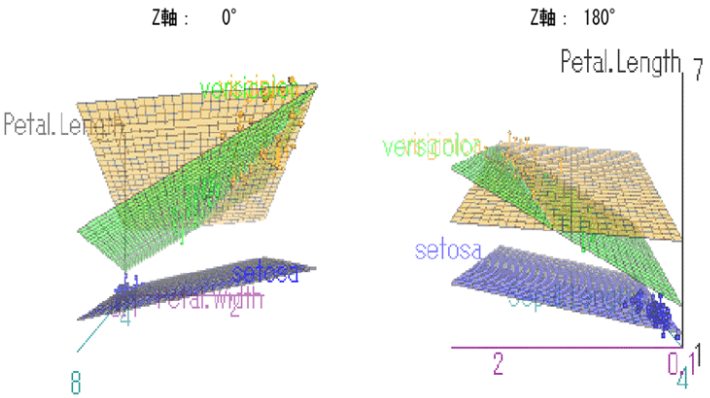
（動画の始まり）　（動画の終わり）

図 2. 分散と不偏分散

5. iris の 3 次元散布図

フィッシャーのアヤメのデータを用いて 3 次元散布図を作成する。R コマンダーの関数を利用し、Petal.Length を目的変数、Petal.Width, Sepal.Length を説明変数として種(setosa, versicolor, virginica) ごとに回帰平面が引かれている。ここで一方向からだと構造が見えにくいので、回転させて全体的な構造が見えるようにした(図 3)。このようにプレゼンテーションの最中に、見えにくい構造の裏側を動画で見せるというテクニックはいかがだろうか。

3 次元図の PNG ファイルへの吐き出しには、通常の PNG 関数は使えず、rgl.snapshot 関数を使わなければならないが、無理なく動画にすることができた。また R コマンダーの関数を用いても問題ない。

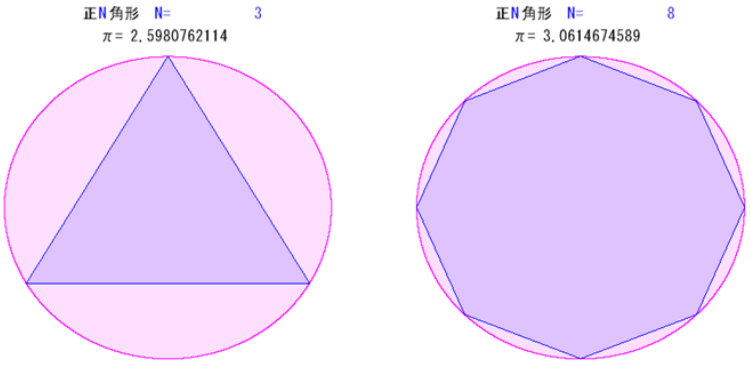


（動画の始まり）　（動画の途中）

図 3. iris の 3 次元散布図

6. 正多角形と円周率

正 N 角形の N 数を増やしていくと円に漸近していく様子を模式的に示した(図 4)。正 N 角形の周の長さから計算される円周率もヘッダー部に表示してある。半径 1 の円に内接する正 N 角形を考えたときの円周率=正 N 角形の周の長さ÷2 が表示されている(正 6 角形ならば円周率=3 と表示される)。N 数を増大させると真の円周率に近づいていくのが分かる。本プログラムのソースコードは巻末の付録に掲載した。

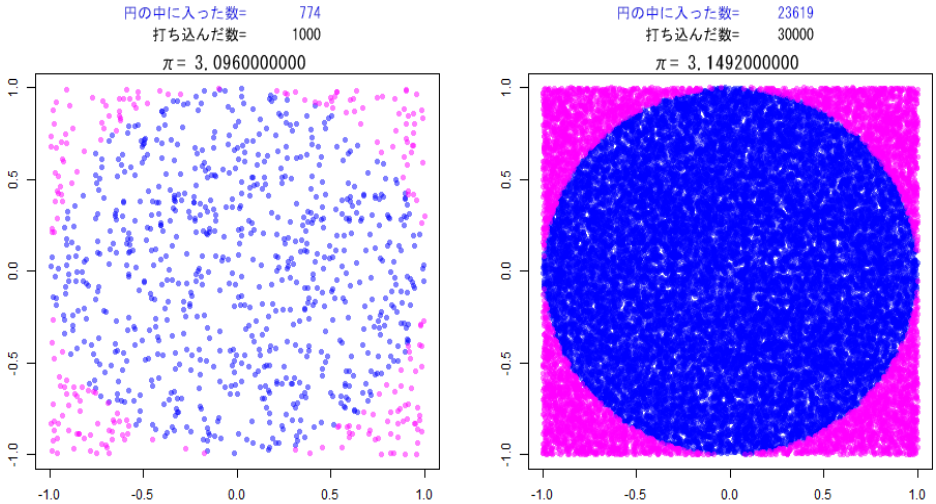


（動画の始まり）　（動画の途中）

図 4. 正多角形と円周率

7. モンテカルロ法による円周率の計算

モンテカルロ法による円周率を計算してみる。面積 4 の正方形の中に 2 次元一様乱数(N 個)を打ち込み、その正方形に内接している半径 1 の円の中に入った個数をカウント(R 個)する。円周率πは 4×R/N に近似できる。すなわち 4 回に 3.14 回くらいは的の中に入るはず。打ち込む乱数の数を増やしていけばπは真の円周率に近づいていくはずである。しかし見て頂くと分かるが、モンテカルロ法で計算された円周率はそれほど精度の高いものではない(図 5)。R ではドットの色を半透明にできるので、散布図を簡単に濃淡図に仕立てることができる。

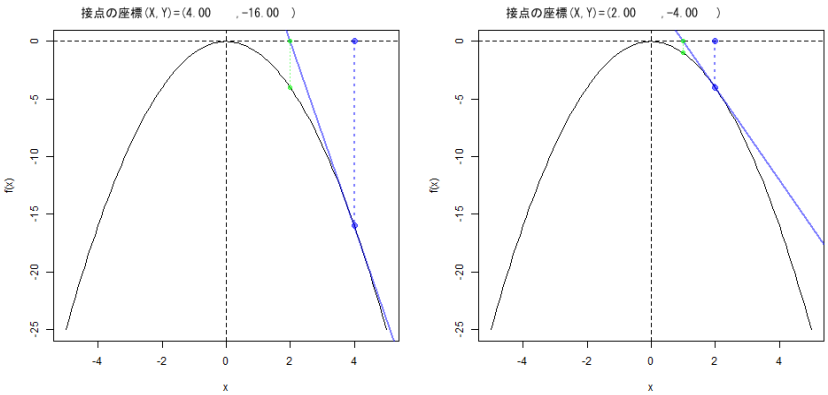


（動画の始まり）　（動画の終わり）

図 5. モンテカルロ法による円周率の計算

8. ニュートン・ラフソン法の概念

単純な 2 次曲線 y=-x^2 を用いて、ニュートン・ラフソン法の概念を説明する(図 6)。先ず最初に x=4 の位置にこの曲線の接線が引かれている(曲線との接点の座標は(4,-16)である)。その接線は x=2 の位置で y=0 と結ばれている。その x=2 に対応した座標(2,-4)が次に接線が引かれるべき座標である。このようにどんどん接線を移動させていくと x 軸方向での青色の点と緑色の点の幅が狭まっていくのが分かると思う。その幅があるしきい値以下に達したとき、収束したとみなしてアルゴリズムを止める。この動画の最後では、この接線が y=0 の線とほぼ一致して止まる。本動画は最尤法を意識して作成したものである。

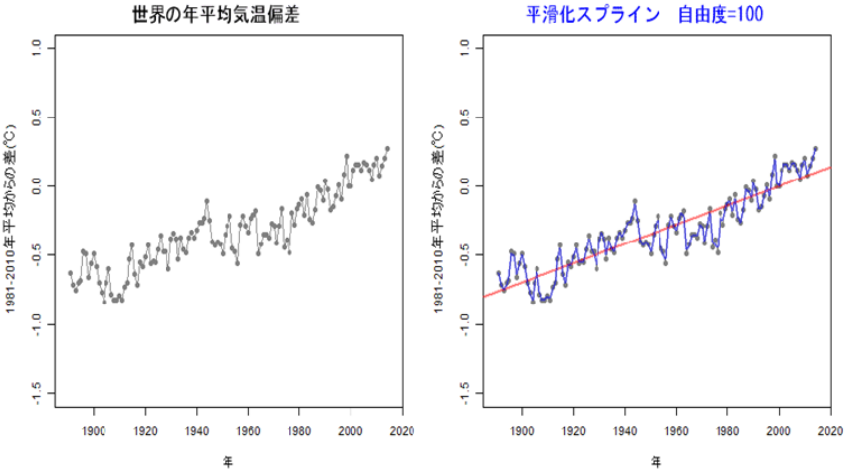


（動画の始まり）　　（動画の途中）

図 6. ニュートン・ラフソン法の概念

9. 線形回帰と平滑化スプライン

気象庁公開の世界の年平均気温偏差のデータ(「http://www.data.jma.go.jp/cpdinfo/temp/list/an_wld.html」)を用いて線形回帰と平滑化スプラインの違いを示す(図 7)。124 年分の年平均気温偏差データであるが、先ずは線形回帰を行ったところ、0.70℃/100 年位のトレンドで世界の気温が上昇していることが分かる。次に GAM プロシジャを用いて平滑化スプラインによる回帰にかけてみた。proc GAM の自由度を増やすごとに"グネグネ度"が増して、実データに沿うようになる。動画の最後では青色の線は実データにほぼ一致してしまう。ここで、線形回帰は精度が悪いがバイアスが小さい、平滑化スプラインは精度が高いがバイアスが大きい手法だと言うことが分かる。

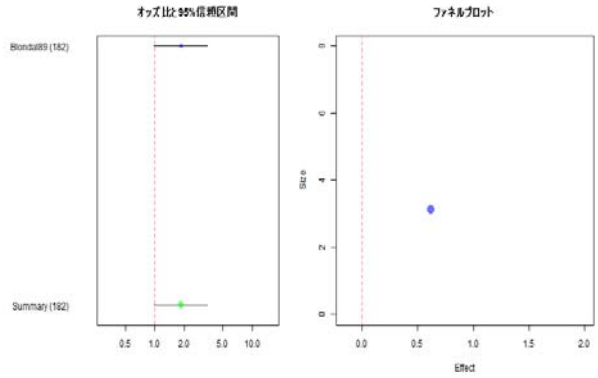


（動画の始まり）　　（動画の終わり）

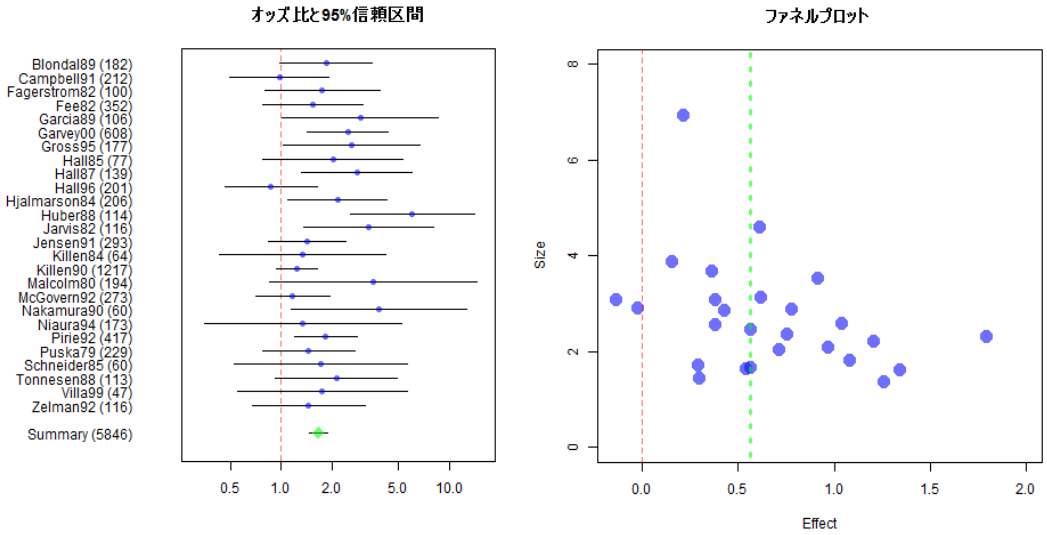
図 7. 線形回帰と平滑化スプライン

10. ニコチンガム研究のメタアナリシス

rmeta パッケージ収録のデータ smoking を用いてメタアナリシスの概念を説明したい[4)]。ニコチンガムを用いたランダム化試験を併合していくというものである。ほとんどの試験で、ガムを処方された人の方が、されなかった人よりも禁煙を達成できたというデータである(2 試験ほどオッズ比が 1 を下回っている)。「オッズ比と 95%信頼区間」(以下左図)と「ファネルプロット」(以下右図)の 2 つのパーツの動画を作成した(図 8)。左図は 1 試験ごとに併合していき、その併合オッズ比(CMH 流オッズ比)が、試験が入るごとに変動している。試験名の次の括弧内に症例数が表示されており、より大規模な試験、あるいは禁煙効果の高かった試験が併合されれば Summary(緑色の点)の変動は大きい。Summary の例数が増えればその 95%信頼区間のヒゲの長さは短くなって行く。左図は freq プロシジャの計算を用いて作成されている。右図は公表バイアスがないかどうか監視している図であるが、rmeta パッケージの関数を用いて作成されている。縦軸の Size とは 1/SE であり、例数の多い試験ほど値が大きい。横軸の Effect とは左図のオッズ比の自然対数をとったものに相当している。緑色の破線は左図の Summary の変動と連動している。本ファネルプロットからは、公表バイアスの強い証拠は見られなかった。



（動画の始まり）

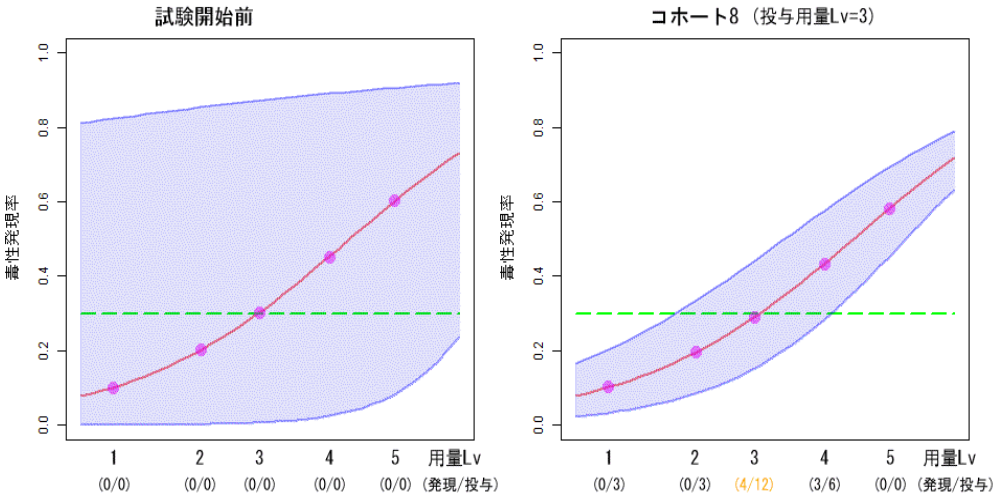


（動画の終わり）

図 8. ニコチンガム研究のメタアナリシス

11. CRM における用量反応曲線の遷移

本章では、矢田ら，日本臨床薬理学会誌 2015 Vol.46 No.1 掲載の「抗がん剤の第 I 相試験における MTD 推定に有用な CRM の歴史的発展と将来の展望」[5)]の Figure6 用量反応曲線の遷移を模式的に示した動画を作成している(図 9)。 1 つのコホートには 3 例の症例がエントリーできる。毒性発現率のターゲットは 0.3 に設定されている。例えばこの動画では、コホート 1 で 3 例エントリーして毒性発現した症例は 0 例であった。その結果用量反応曲線は変形し、ターゲット 0.3 に近い用量レベルは 4 番目の用量であるが、用量レベルは次のコホートに移るときに 1 段階しか上げられないルールなので、コホート 2 では 2 番目の用量レベルが設定される。例えばコホート 5 まで行ったときに、用量レベル 4 で累積で 6 例中 3 例が毒性発現して、用量反応曲線が変化し、ターゲット 0.3 に近い用量レベル 3 に次コホートで下げられるというイメージである。動画を眺めていくと例数の集積に伴って 90%信用区間の幅が狭まっていくのが見て取れる。



（動画の始まり）　　（動画の終わり）

図 9. CRM における用量反応曲線の遷移

12. まとめ

動画にすることによって、静止画だけでは説明しにくい複雑な状況を説明することができる。また R のグラフの活用によってより美しい動画の作成が可能となった。動画作成の一連のプロセスは SAS の一つのプログラムで可能である。この手法によって R 独自関数を SAS のプログラムに織り交ぜることが可能である。実際に本論文に使用したプログラムはほぼ R しか使用していないものも存在するし、多くを SAS に頼ったものもある。本論文をきっかけとして、多くの SAS ユーザーが少しでも R を意識して頂けたら幸いである。

参考文献

1) 関根 暁史 (2014). 動画による統計表現 ～新しい統計の要約～，SAS ユーザー総会
2) 長谷川 要 (2002). スピログラフを再現しよう －GIFANIM Device Driver を用いたアニメーション図形の作成－，日本 SAS ユーザー会
3) 舟尾 暢男 (2009). SAS でベイズ推定を行う方法 －proc MCMC vs R&WinBUGS－，SAS ユーザー総会
4) 大門 貴志 (2010). R による統計解析ハンドブック -第 2 版-，(株)メディカル・パブリケーションズ
5) 矢田 真城 (2015). 抗がん剤の第 I 相試験における MTD 推定に有用な CRM の歴史的発展と将来の展望，日本臨床薬理学会 2015 Vol.46 No.1

付録

```
/* 図 4 プログラム */
---------------------------------------------------------------------------
NOTE1            :  プログラム中の以下の URL は書き換えてご使用下さい。
「C:¥Users¥Documents¥」
NOTE2            :  プログラム中の以下のバージョンは書き換えてご使用下さい。
「R-3.1.1」
---------------------------------------------------------------------------;

*以下のパスを書き換えてご使用下さい;
%let _path=C:/Users/Documents;
*;
%annomac;

/* 円に見立てる */
data DEF;
   do i=1 to 500000;
      output;
   end;
data CIRCLE; set DEF;
   theta=i*360/500000;
   X1=sin(theta*constant('pi')/180);
   Y1=cos(theta*constant('pi')/180);
run;
libname doc xport "C:¥Users¥Documents¥TMP1.xpt";
proc copy in=work out=doc;
   select CIRCLE / memtype=data;
run;

%macro onestate(state);
data PAI;
   pai=&state.*cos((90-180/&state.)*constant('pi')/180);
run;
/* ヘッダー部の注釈 */
data ANNO2; set PAI;
   %DCLANNO;
   %SYSTEM( 3, 3, 3 );
   length text $80;
```

```
   text="正";
   %label(35,95,text,black,0,0,3.50,'ＭＳ　ゴシック',5);
   text="N";
   %label(38,95,text,blue,0,0,3.50,'ＭＳ　ゴシック',5);
   text="角形";
   %label(43,95,text,black,0,0,3.50,'ＭＳ　ゴシック',5);
   text="N=";
   %label(52,95,text,blue,0,0,3.50,'ＭＳ　ゴシック',5);
   text=compress(put(&state.,best.));
   %label(75,95,text,blue,0,0,3.50,'ＭＳ　ゴシック',4);
   text="π=";
   %label(40,90,text,black,0,0,3.50,'ＭＳ　ゴシック',5);
   text=compress(put(round(pai,.0000000001),32.10));
   %label(67,90,text,black,0,0,3.50,'ＭＳ　ゴシック',4);
run;
data DEF;
   do i=1 to &state.;
      output;
   end;
data TEMP; set DEF;
   theta=i*360/&state.;
   X1=sin(theta*constant('pi')/180);
   Y1=cos(theta*constant('pi')/180);
run;
libname doc xport "C:¥Users¥Documents¥TMP2.xpt";
proc copy in=work out=doc;
   select TEMP / memtype=data;
run;
filename TEXT "C:¥Users¥Documents¥TMP.R";
data _NULL_;
   file TEXT;
put "library(foreign)";
put "x <- read.xport('&_path./TMP1.xpt');y <- read.xport('&_path./TMP2.xpt'); ";
put "png('&_path./WRK.png');";
put "plot.new();plot.window(xlim=c(-1, 1), ylim=c(-1, 1));";
put "polygon(x$X1,x$Y1,col='#ff00ff20',border='#ff00ff');";
put "polygon(y$X1,y$Y1,col='#0000ff20',border='#0000ff');";
put "dev.off()";
```

```
run;

options noxwait ;
X "'C:¥Program Files¥R¥R-3.1.1¥bin¥R.exe"
  --no-save < "C:¥Users¥Documents¥TMP.R" > "C:¥Users¥Documents¥TMP.log'";

data ANNO1;
    length function $8 style $ 32 ;
    retain xsys ysys '3' hsys '3' when 'a';
    function='move'; x=0; y=0; output;
    function='image'; x=100; y=100;
    imgpath="C:¥Users¥Documents¥WRK.png"; style='fit'; output;
run;
data ANNO; set ANNO1 ANNO2; when="a"; run;
proc ganno anno=ANNO;
run;
%mend;

data usa;
    do
state=3,4,5,6,7,8,12,24,48,96,192,384,768,1536,3072,6144,12288,24576,49152,98304,196608,393216;output;end;
run;

data _null_;
      set usa end=done;
      file 'C:¥Users¥Documents¥plot.sas';
      if _n_ = 1
      then put "filename animmap 'C:¥Users¥Documents¥図 4_円周率.gif';" /
                    "goptions reset=goptions    device=gifanim gsfmode=replace gsfname=animmap "
                    "cback=white iteration=0 delay=250 disposal=background noborder htitle=13pt xmax=5in ymax=5in;";
      else if _n_ = 2
      then put "goptions gsfmode=append;";
      if done then put "goptions gepilog='3B'x delay=1000;";
      put '%onestate(' state ');';
run;

%inc 'C:¥Users¥Documents¥plot.sas';
```