動画による統計表現

～新しい統計の要約～

関根　暁史
株式会社 ACRONET／生物統計部

Dynamic statistical graphs

Satoshi Sekine
ACRONET Corp./Biostatistics Dept. Data Science Division

要旨

SAS には動画を簡単に作成できる機能がある。数式に依存しない形で、多くの人に統計の概念を伝えることができるであろう動画の試作品を作成したので紹介する。

キーワード：SAS グラフ，GIF

1. はじめに

SAS には、複数枚の SAS グラフをコマ送りにして動画としてしまう機能が備わっている [1)2)]。この機能を活かし、ほとんど数学を使わない動画の形で統計を表現してみたら、統計初心者にも統計学の根本が伝わると考えた。本論文では「動的な三次元図」「動的な分割表」「動的な・・・」という章立てとして、統計学にも様々な分野があるが、分野にとらわれることなく、それぞれの章に最適と思われるテーマを用意し、そのテーマの説明補助となり得るような動画の見本を試作した。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. 動画作成プログラム

図 0. SAS 社 HP 掲載のプログラム

SAS 社ホームページに図 0 作成のプログラムが掲載されている「http://support.sas.com/kb/25/255.html」。図 0 はアメリカ合衆国の地図における州の色が経時的に塗られていくというものである。同プログラムは拡張子 GIF となる動画を作成して吐き出す。この GIF ファイルはパワーポイントに貼りつければ、スライドショーにすると駆動するので、プレゼンテーションの最中に動画を見せることが可能である。またプレゼンテーションを行う環境は SAS がインストー

ルされている必要もない。同プログラムを小加工して流用するだけで、SAS グラフが描ける人ならば誰でも簡単に動画が作成可能である。よって以降の動画は図 0 の作成プログラムを元に作成することとする。

3. 動的な三次元図

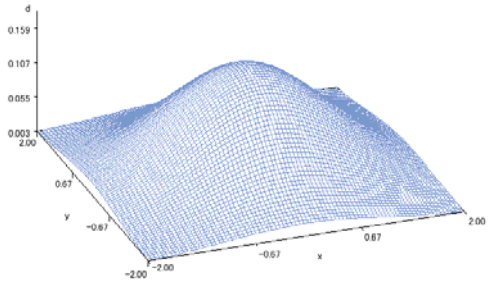


図 1. 二次元正規分布

「SAS によるデータ解析入門」(東京大学出版会)[3]p.135～に掲載のデータ normal は相関係数 r=0.6 の時の 2 次元正規分布を示している。図 0 のプログラムを利用して、この相関係数 r を 0 から 0.9 まで 0.1 ずつ変化させた三次元図を作成することを考える。図 0 のプログラムで回転していたのは&state というマクロ変数のみであったので、相関係数 r=&state.とおいて、state=0 to 0.9 by 0.1 となるデータセット usa を用意して、図 0 のプログラムをそのまま実行すれば図 1 が完成する。本プログラムのソースコードは巻末のプログラム 1 に掲載した。

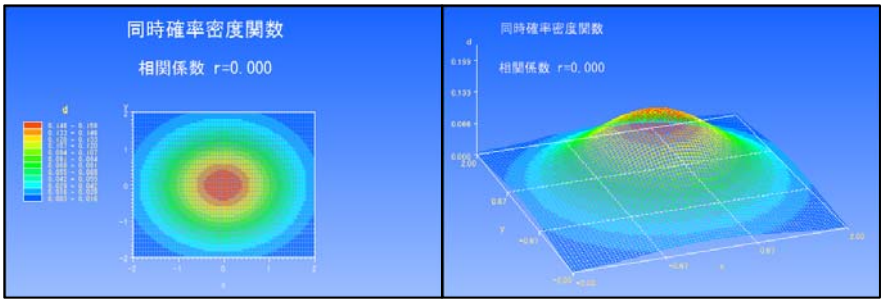


図 2. 二次元正規分布（高品位）

さらに参考文献 4)を参照頂くことで、図 2 のように高品位にすることも可能である。図 2 では 2 枚の別々の図が同時に動いているが、コマ送りの速度を同じにしているので同期して動いて見える。コマ送りの速度は、delay=の値を SAS 側から設定すればよい。パワーポイントにおいて、同時に複数枚の動画を動かいたい際は、ディレイタイムを同じとすることで、無理に 1 枚絵に仕立てる必要はない。本章では確率密度と相関との関係を示す動画として紹介した。

4. 動的な分割表

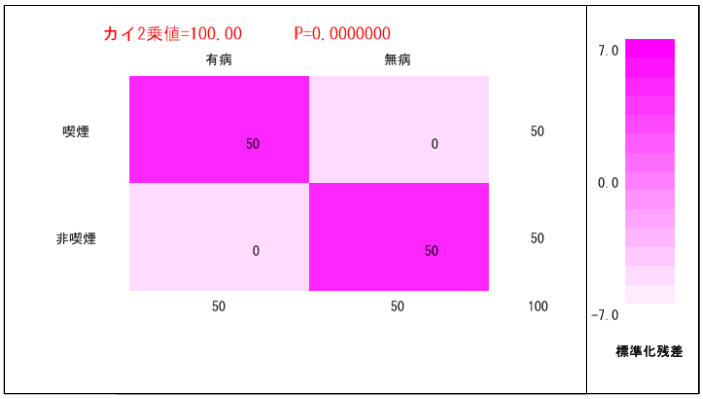


図 3. カイ 2 乗検定

度数とともに色の濃淡が変化する分割表を考えた。（喫煙・非喫煙）×（有病・無病）の 2×2 分表（人工データ）であるが、周辺分布を固定しながら各セルを 1 例ずつ変化させていく。セル色の濃淡は標準化残差の値と紐付いている。すなわち色の濃いセルには度数が集中しているし、色の薄いセルは期待値と比較して度数の少ないことが判る。全てのセルの度数が期待値と変わらない場合は、一様の平面が出来る。この一様性が崩れるほどカイ 2 乗値が跳ね上がる。P 値は 5%有意の時、赤字で表現されるが、喫煙・無病が多い方向性で 5%有意になる際は、緑色となる。

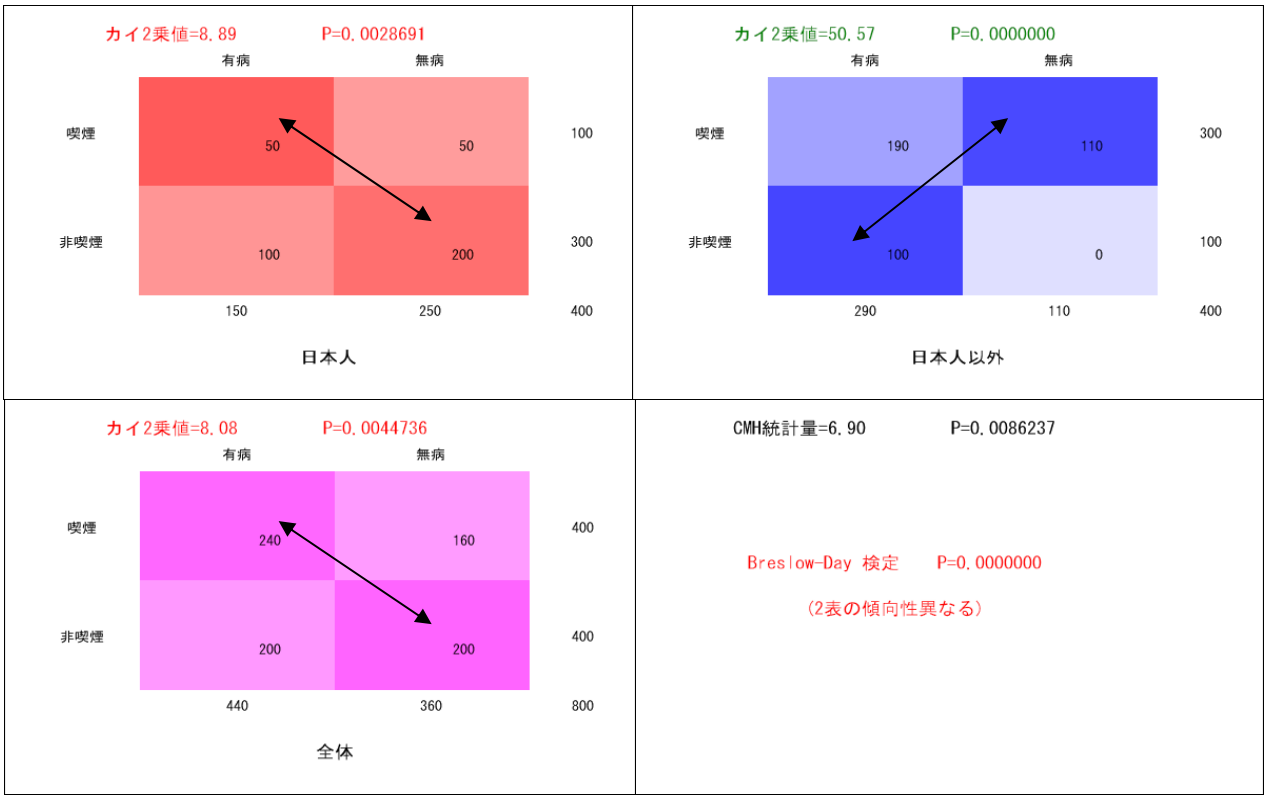
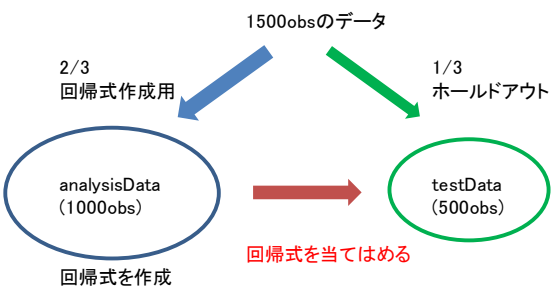


図 4. 層別カイ 2 乗検定

図 3 の概念の分割表を 2 枚同時に動かすことを考える。日本人と日本人以外のそれぞれ 400 例ずつの分割表のセル度数は変化させるが、全体（日本人+日本人以外）800 例のセル度数は全く変化させないようにする。この図 4 の状態で、全体と日本人では 5%有意であるが、日本人以外は喫煙・無病の方向性で 5%有意の状態である。表の傾向性はセル色の濃い部分をたどれば見えてきて、この図中には傾向性の矢印を書き込んでいる。図 4 は、日本人と日本人以外は別の傾向性を持っているにもかかわらず、全体の P 値が有意となっている。これは右下の Breslow-Day 検定にも反映されていて、Breslow-Day 検定が 5%有意の時“(2 表の傾向性異なる)”と赤字で表示するようにした。すなわち当該図表はシンプソンのパラドクスを示したものである。日本人と日本人以外が、最後まで全体と同じ傾向のまま有意にならないという人工的データを作成して本動画を作成した。つまり全体の有意をもたらしているのは、喫煙か非喫煙がではなく、国別という原因が作用しているのではないかという例を示した。

5. 動的な折れ線



1,500obs のデータを analysisData（3 分の 2）と testData（3 分の 1）に分割し、analysis 側で線形重回帰式を作成してその回帰式を test 側に適用することを考える。変数選択を伴う回帰分析において、analysis 側、test 側とともに ASE（残差平方和を N 数で割ったもの）を計算して逐次お互いの ASE を

比較する。analysis 側は最小 2 乗法によって ASE を減らしていくが、test 側は言わば受身的に ASE を計算させられることになる。本データは SAS ヘルプの GLMSELECT の章にあるものを用いているが(Example 42.2 Using Validation and Cross Validation)、本回帰分析は下記のソースコードの通りに行った。

```
proc glmselect data=analysisData testdata=testData;
    class c1 c2 c3(order=data);
    model y =   c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
              / selection=stepwise(select = sl)
                hierarchy=single;
run;
```

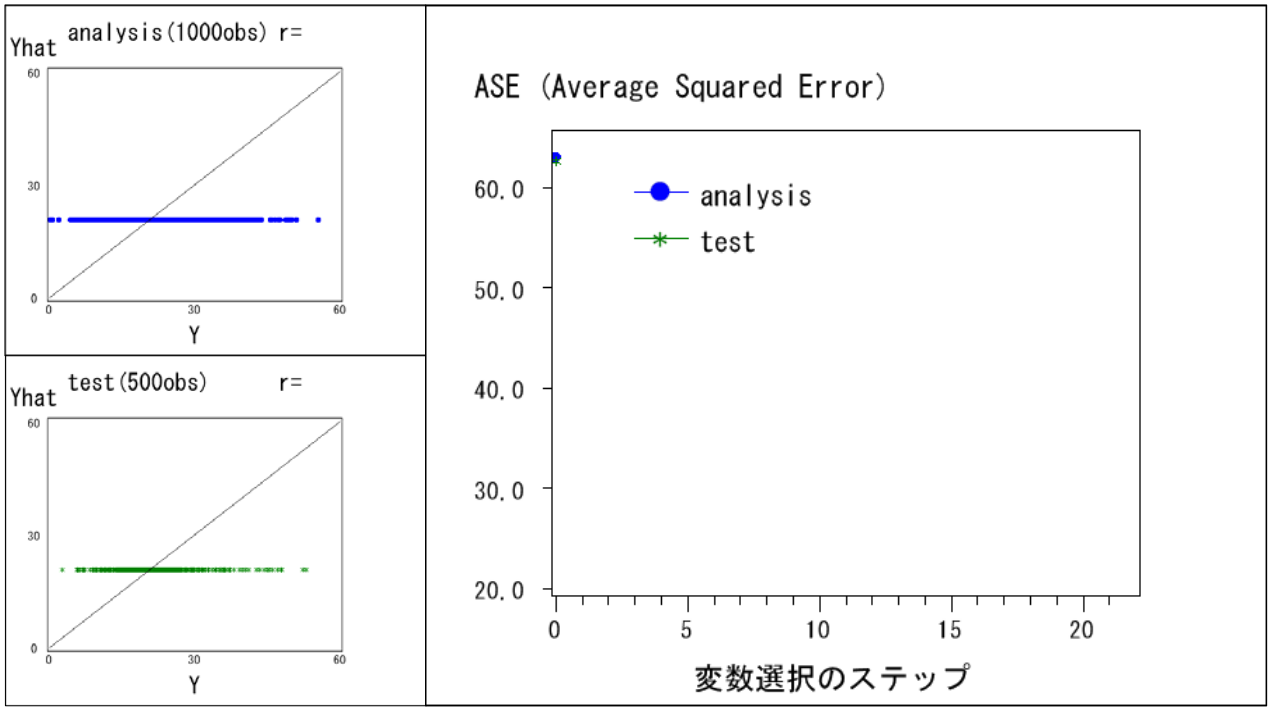


図 5. 線形重回帰分析における過学習の概念

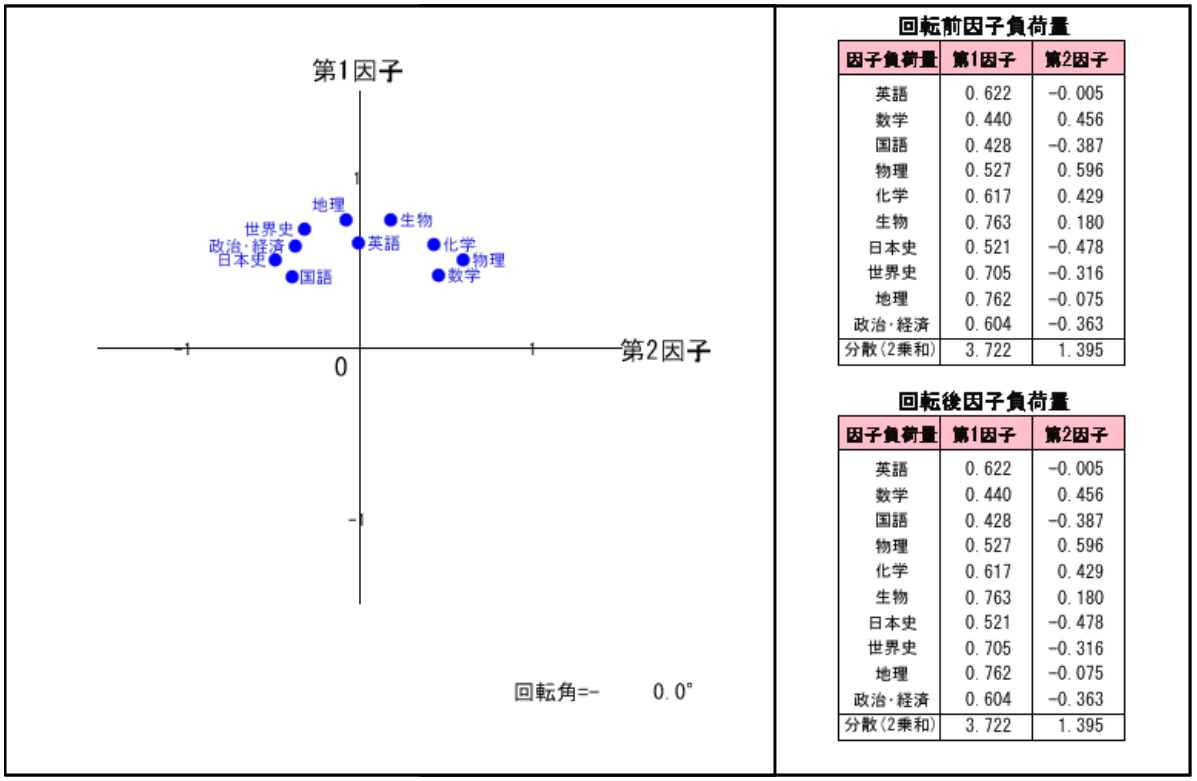
上記プログラムを実行した時の ASE の変動の過程を見たものが図 5 である。左に補助的に Y と Yhat の散布図を相関係数とともに付けている。変数選択のステップが進むごとに analysis 側、test 側ともに ASE は下がり、散布図の分布は 45° の対角線に近づいていく（相関係数は上昇していく）。しかし 10 ステップ目で test 側は ASE が最小(相関係数は 0.790)になった後、11 ステップ以降 ASE は上昇していくことになる。よって analysis 側は過学習をしていることが考えられ、analysis 側のステップは 10 ステップ目付近で止めておくことがバイアス減少のために相応しいと思われる。本動画は回帰分析における過学習の概念を伝えるものであり、11 ステップ目以降になると“Over Learning”と赤字で表示するようにしている。

6. 動的な座標軸

10 教科 50 人分の人工データを用意して因子分析を行って見る。下記教科データは「SAS によるデータ解析入門」[3)]p.193 掲載の認知課題データを小加工したものである。

英語	数学	国語	物理	化学	生物	日本史	世界史	地理	政治･経済
63	57	74	56	25	63	66	68	88	78
66	40	83	56	65	70	62	72	76	60
･･･									
60	52	78	52	80	63	70	54	76	52

本データを nfact=2 の FACTOR プロシジャに供する。プロシジャのデフォルトのまま主成分解を解くものとする。第 1 因子を縦軸、第 2 因子を横軸として因子負荷量の散布図を描く(図 6)。既に因子軸の回転前から第 1 因子は 10 教科の総合得点を、第 2 因子はいわゆる文理を意味していることが想像できる。回転前因子負荷量の分散(2 乗和)は、(第 1 因子，第 2 因子)=(3.722，1.395)であった。この因子軸をバリマックス法によって直交回転させてみる。



回転前因子負荷量

因子負荷量	第1因子	第2因子
英語	0.622	-0.005
数学	0.440	0.456
国語	0.428	-0.387
物理	0.527	0.596
化学	0.617	0.429
生物	0.763	0.180
日本史	0.521	-0.478
世界史	0.705	-0.316
地理	0.762	-0.075
政治･経済	0.604	-0.363
分散(2乗和)	3.722	1.395

回転後因子負荷量

因子負荷量	第1因子	第2因子
英語	0.622	-0.005
数学	0.440	0.456
国語	0.428	-0.387
物理	0.527	0.596
化学	0.617	0.429
生物	0.763	0.180
日本史	0.521	-0.478
世界史	0.705	-0.316
地理	0.762	-0.075
政治･経済	0.604	-0.363
分散(2乗和)	3.722	1.395

図 6. 因子分析における直交回転の概念

動画では-45° 方向に座標軸を回すことによってバリマックス回転をイメージした。バリマックス回転後、(第 1 因子，第 2 因子)=(2.687，2.429)となり分散がより平均化した。「数学」「物理」などに着目すると、回転後因子負荷量の第 1 因子にはほとんど寄与しなくなった。逆に「国語」「日本史」などに着目して見ると第 2 因子には寄与しなくなって、単純構造が得られていることが判る。第 1 因子は“文系能力”を、第 2 因子は“理系能力”を表していると解釈できるので回転の終わりに赤字で表示した。本動画は、回転前・回転後で座標軸の直交性が崩れていないということと、各教科間の因子負荷量の内積(相関性)に変化が無いということを示すために作成した。

7. 動的なデンドログラム

6 章の人工データをそのまま用いて、変数のクラスター分析を行うこととする。デフォルトの VARCLUS プロシジャ [5]によりクラスター数を上昇させていく実験を行う。初期状態の分割クラスター数が 1 の時の分散説明率 0.372 とは、6 章の第 1 因子の分散 3.722（すなわち主成分分析の第 1 固有値）を教科数 10 で割った値と一致する。分割クラスター数を 2 とした時、文系的な第 1 クラスター(英・国・日・世・地・政)と理系的な第 2 クラスター(数・物・化・生)に割り付けられる。因子分析のオーソブリク回転によって第 1 次割付がなされる。回転後因子負荷量の絶対値が第 1 因子の方により寄与していた教科は第 1 クラスターに割り付けられ、第 2 因子の方により寄与していた教科は第 2 クラスターに割り付けられたのである。実際のアルゴリズムは k-means クラスタリングによく類似していて、因子分析のオーソブリク回転以降に、主成分分析による第 2 次割付へと反復されるのであるが､本データでは maxirer=1 で全て十分収束してしまうので反復の詳しい解説は割愛する｡分割クラスター数=2 において､第 1 クラスターの中での主成分分析の第 1 固有値は 2.739、第 2 クラスターの中での主成分分析の第 1 固有値は 2.153 であるので、合計値 4.892 を 10 で割ったものが分散説明率となっており、クラスター数 1→2 の上昇で、説明率 0.372→0.489 に上昇した。

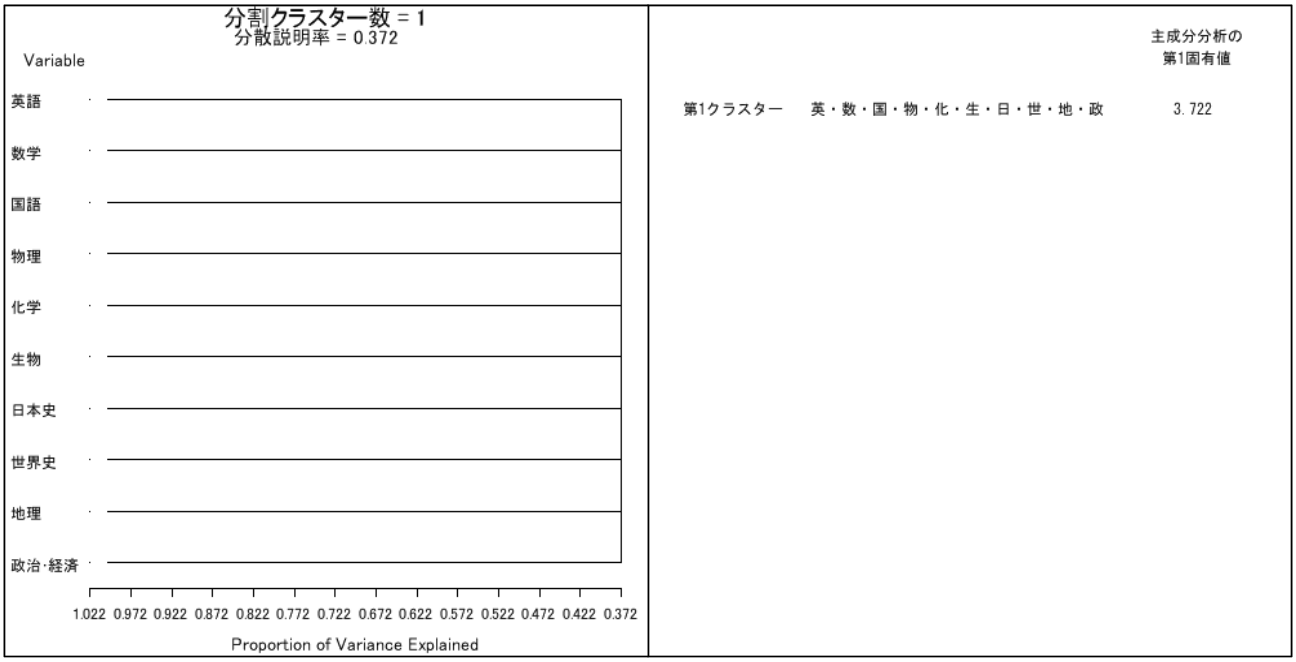


図 7. 変数の階層的クラスター分析

その後動画を見ていくと、理系の中でも「数物クラスター」･「生化学クラスター」に割れたり、文系の中でも「言語クラスター(英・国)」が現れたりする。10 教科しかないので 10 個のクラスターまで分割して、分散説明率が元の 1 となって終了である（固有値もそれぞれ 1 ずつとなって終わる)。本動画は、VARCLUS プロシジャがクラスター分析とは言っても主成分・因子分析に近い考え方をしていること紹介するために作成して見た。

8. 動的なしきい

ある臨床検査薬を考える。500 例の有病群は正規分布 $N(70,15^2)$に従っており、9,500 例の無病群は正規分布 $N(40,15^2)$に従っている(すなわち有病率 5%)。しきい値を超えた場合を陽性(+)とみなし、それ以外は陰性(-)である。しきい値を 30 から上昇させたとき、感度と特異度の変化を図 8 に表した。しきい値は ROC 曲線上も動いている。

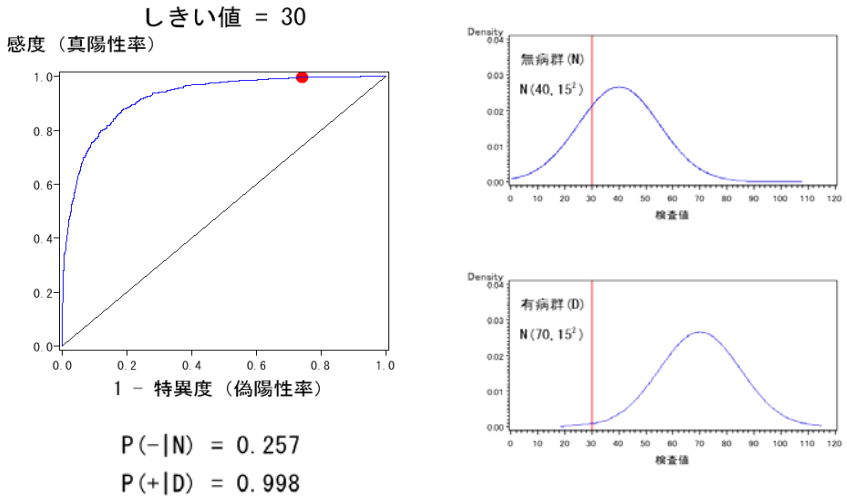


図 8. ROC 曲線としきい値

9. 動的な ROC

500 例の有病群は正規分布に従っており、標準偏差は 15 のまま平均が 55 から 80 に変化させる。9,500 例の無病群は正規分布に従っており、標準偏差は 15 のまま平均が 55 から 30 に変化させる(2 群は等分散としている)。群間差が開くに従って ROC 曲線の AUC が 0.5 から上昇する様子を図 9 に示した。t-検定にも考え方が近いので群間差の t-統計量の表示も添えた。

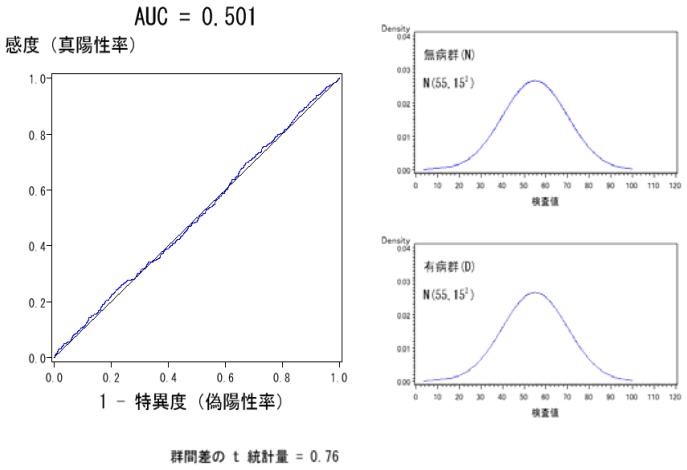
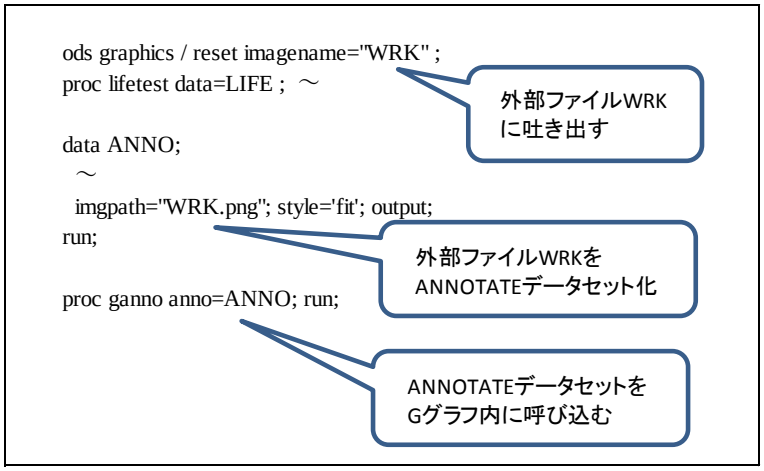


図 9. AUC と群間差の関係

10. 動的なカプランマイヤー図

本章ではカプランマイヤー図を動かすということだけではなく、SG グラフを動画にするということを同時に試みている。goption である GIFANIM Device Driver は残念ながら SG グラフ(すなわち ods graphics on にて出力されるグラフ)はサポートしていない。しかし一工夫することで SG グラフを動画にすることができる[6]。本プログラムのソースコードは巻末のプログラム 2 に掲載した(グラフ中の検定統計量表示部分は、紙面の関係上割愛させて頂いた)が、SG グラフを動画化する手順を以下に記す。



手順 1：ods graphics 機能によって SG グラフを外部ファイル(拡張子 png)として吐き出す

手順 2：外部ファイルをそのまま ANNOTATE データセット化してしまう

手順 3：ANNOTATE データセットを ganno プロシジャに呼び込んで回転させ、あたかも SG プロシジャが回転していることにしてしまう

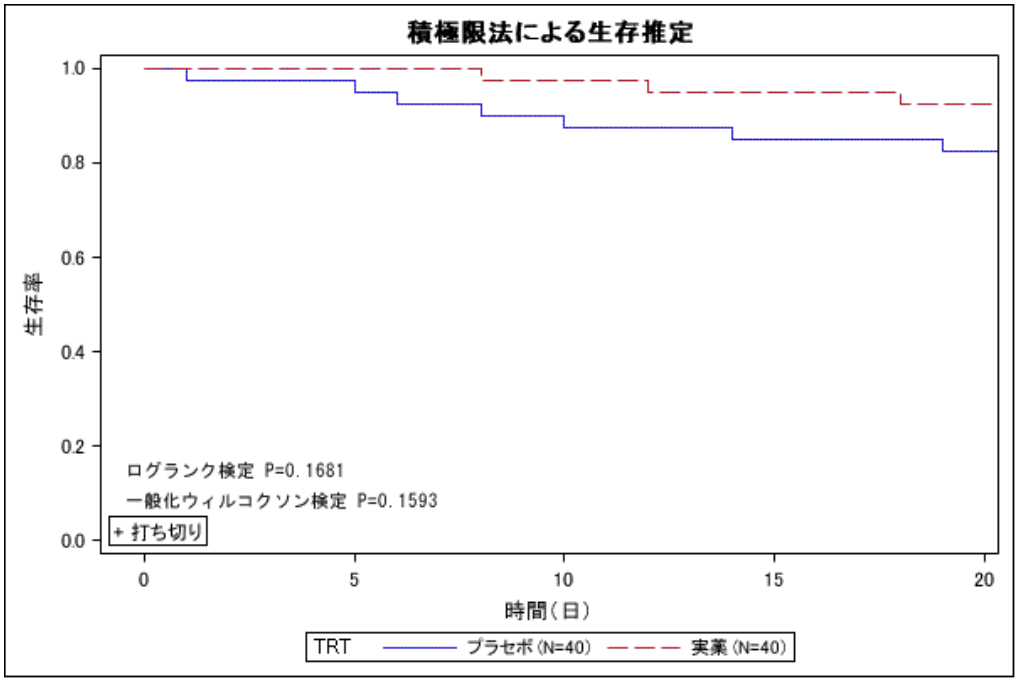


図 10. ログランク検定とウィルコクソン検定の比較

図 10 は、プラセボ群と実薬群の生存率を 20～600 日まで追跡した時の、ログランク検定と一般化ウィルコクソン検定の違いを示した動画である。最初は実薬群の生存率が勝っているように見えるが、300 日付近で生存率が逆転するという人工データを作った。一般化ウィルコクソン検定は比較的初期の差を見ているのに対し、ログランク検定は時間軸の後方の差を見ているのがお判りいただけるとと思う[7]。

11. 動的なクラスター

有名なフィッシャーのアヤメのデータ(1936)を用いて、k-means クラスタリングの概念を動画にすることとした(図 11)。データには 4 変数あるが、二次元の散布図で表現したいため、そのうち花弁の幅・花弁の長さのみを用いることとする。本 150 件のデータを FASTCLUS プロシジャによって 3 分割する。反復数を少なくしたいため、初期シードはそれぞれのクラスターに近いデータを代表値とした。左にはオリジナルの種(セトサ・バーシカラー・バージニカ)ごとの散布図を参考までに置いた。

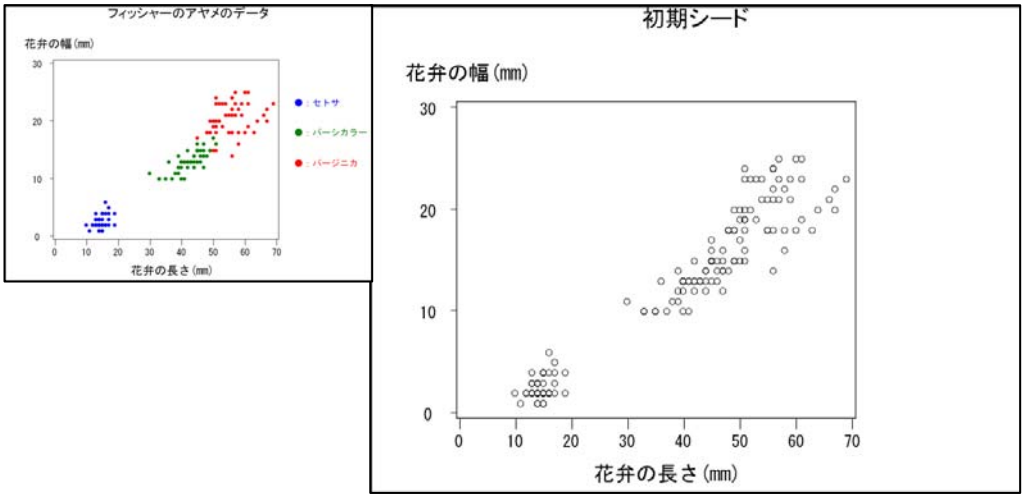


図 11. k-means クラスタリングの概念

動画では先ず初期シード（代表値）が設定され、それぞれ 3 つのシードからユークリッド距離が近い点がそれぞれ 3 つのクラスターに割り付けられる。それぞれのクラスターの重心が計算され、その重心を第 2 のシードとしてクラスタリングが繰り返される。前回の割付と矛盾がなかった場合、収束したとみなし、クラスタリングは終了する。この 1 回の実験は計 4 回の反復で終了し、150 件のうち 8 件が左図と比較して誤答であったが最終クラスターはオリジナルのデータに近い分類となった。

12. まとめ

動画は大量の情報を 1 枚に集約して表現することができる。プレゼンテーションにおいて結果だけでなくプロセスを説明するのに適している。複数枚のタイプの異なる動画を同期的に動かすことで、統計量の多面的な連動を表現することができる。動画とすることで何かシミュレーション（実験的なこと）を行っていることを伝えることができ、説明を簡略化できる。上記の動画は全て SAS9.2 を用いて作成した(※)。SAS で動画を作成することは特殊な外部オプションを必要とせず容易なことなので、ご自身のニーズに合った動画作成にチャレンジして頂きたい。

(※)SAS9.3 以上で実行される場合は、メイン SAS ウィンドウの上部にあるメニューからツール→オプション→プリファレンスを選択、結果タブを表示し、「HTML を作成する」のチェックを外し、「リストを作成する」にチェックを入れて実行すると上手く行くでしょう。

参考文献

1) 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
2) 岸本 容司 (2003). SAS グラフによる動く万華鏡の作成, 日本 SAS ユーザー会
3) 竹内 啓 (1994). SAS によるデータ解析入門[第 2 版], 東京大学出版会
4) 関根 暁史 (2012). 色を自在に操る（HSV カラーコードのすすめ), SAS ユーザー総会
5) 岸本 淳司 (1996). 変数のクラスタリング－PROC VARCLUS 再発見－, 日本 SAS ユーザー会
6) Xin Zhang (2013). Extended SAS GIFANIM Device Usage on Table Reporting and Template-Based Graphics, SAS Global Forum
7) 大橋 靖雄 (1995). 生存時間解析 SAS による生物統計, 東京大学出版会

付録

```
/* プログラム 1 */
%macro onestate( state, ds );
data normal;
   r=&state.; pai=3.141593;
   c=(1/(2*pai*(1-r**2)**0.5));
   do x=-2 to 2 by 0.05;
        do y=-2 to 2 by 0.05;
            d=c*exp(-(0.5/(1-r**2)*(x**2-2*r*x*y+y**2)));
            output;
        end;
    end;
run;
proc g3d data=normal;
    plot y*x=d/rotate=20 tilt=40;
run; quit;
%mend;

data usa;
   do state=0 to 0.9 by 0.1; output; end;
run;

data _null_;
    set usa end=done;
    file '～URL 指定～¥normal.sas';
    if _n_ = 1
    then put "filename animmap '～URL 指定～¥NORMAL.gif';" /
                "goptions reset=goptions device=gifanim gsfmode=replace gsfname=animmap xpixels=600
ypixels=400"
                "cback=white iteration=0 delay=150 disposal=background noborder htitle=13pt;";
    else if _n_ = 2
    then put "goptions gsfmode=append;";
    if done then put "goptions gepilog='3B'x;";
    put '%onestate(' state ', usa );';
run;

%inc '～URL 指定～¥normal.sas';
```

```
/* プログラム 2 */
proc format;
   value trt 1="実薬  (N=40)" 2="プラセボ  (N=40)";
run;

data LIFE;
   input TRT TIME CENSOR @@;
   format TRT trt.;
   label TRT="治療" TIME="時間（日）" CENSOR="打ち切り";
   cards;
   1 8     0      2 1     0
   1 12    0      2 5     0
   1 18    0      2 6     0
   1 24    0      2 8     0
   1 36    0      2 10    0
   1 48    0      2 14    0
   1 68    0      2 19    0
   1 84    0      2 22    0
   1 95    0      2 32    0
   1 102 0       2 34    0
   1 109 0       2 40    0
   1 118 0       2 46    0
   1 132 0       2 50    0
   1 144 0       2 54    0
   1 156 0       2 62    0
   1 168 0       2 64    0
   1 174 0       2 66    0
   1 184 0       2 68    0
   1 192 0       2 72    0
   1 198 0       2 74    0
   1 219 0       2 80    1
   1 220 1       2 82    0
   1 232 0       2 86    0
   1 244 0       2 96    0
   1 252 0       2 105 0
   1 264 0       2 120 0
   1 270 0       2 160 0
   1 290 0       2 280 0
```

```
   1 300 0       2 300 0
   1 320 0       2 401 0
   1 333 0       2 501 0
   1 350 0       2 600 0
   1 364 0       2 610 0
   1 400 0       2 650 0
   1 464 0       2 678 0
   1 502 0       2 694 0
   1 555 0       2 700 0
   1 559 0       2 701 0
   1 601 0       2 800 0
   1 602 0       2 900 0
   ;
run;

%macro onestate( state );
ods graphics / reset width=6in height=4in imagename="WRK" ;
proc lifetest data=LIFE maxtime=&state.;
   time TIME*CENSOR(1);
   strata TRT;
run;

data ANNO;
   length function style $ 32 ;
   retain xsys ysys '3' hsys '3' when 'a';
   function='move'; x=0; y=0; output;
   function='image'; x=100; y=100;
   imgpath="WRK.png"; style='fit'; output;
run;

proc ganno anno=ANNO; run;
%mend;

data usa;
   do state=20 to 600 by 20; output; end;
run;
```

```
data _null_;
    set usa end=done;
    file '～URL 指定～¥KM.sas';
    if _n_ = 1
    then put "filename animmap '～URL 指定～¥カプランマイヤー.gif';" /
                "goptions reset=goptions device=gifanim gsfmode=replace gsfname=animmap xpixels=601
ypixels=401"
                "cback=white iteration=0 delay=200 disposal=background border htitle=13pt ;";
    else if _n_ = 2
    then put "goptions gsfmode=append;";
    if done then put "goptions gepilog='3B'x;";
    put '%onestate(' state ');';
run;

%inc '～URL 指定～¥KM.sas';
```