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

※1：スライドショーにしてご覧ください
※2：Windows XP環境ではご覧になれないことがございます

はじめに

- SASでの動画作成は簡単
  - 複数枚のグラフをまとめて1つのGIFファイルとするだけ
  - 動かす環境はSASがインストールされていなくてよい

はじめに

- 動画を利用した初心者でも判り易い
  統計資料を作成
- SAS社HP掲載のプログラムを参照
  - 以下アメリカ合衆国の地図のプログラムを利用
    「http://support.sas.com/kb/25/255.html」

動的な三次元図

- 二次元正規分布を動かす
  - 動的な変数をマクロ変数化

（例）相関係数 r = &state.

  - 動的変数の範囲を定めたデータセットを用意

（例）data usa ;

state = 0 to 0.9 by 0.1 ; output ; end ;

run ;

「SASによるデータ解析入門(東京大学出版会) p.135～」

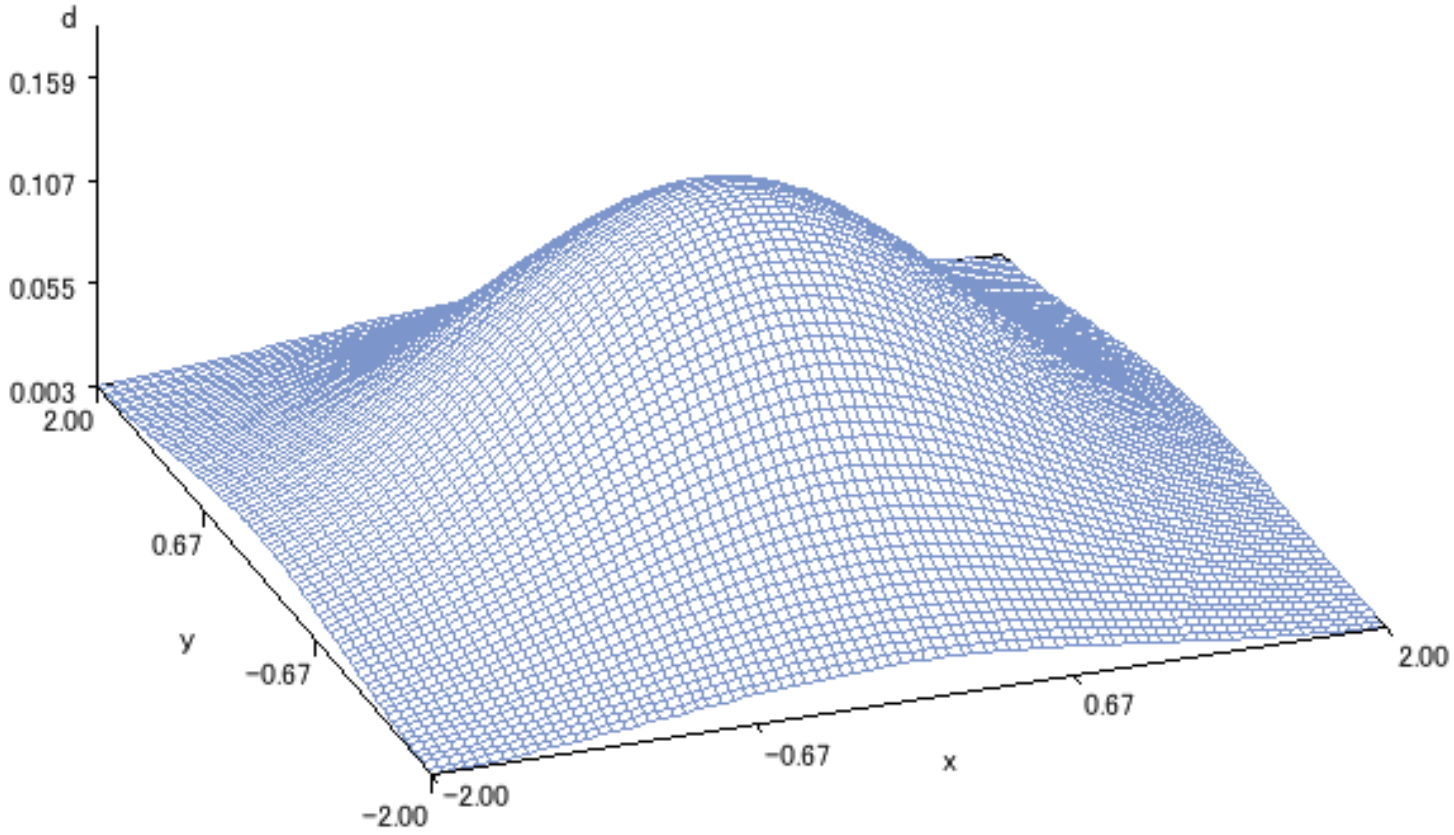


図1. 二次元正規分布

動的な三次元図(2)

- カラフルにする
  - 「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2012(関根)」参照
- コマ送りスピードの調節
  - SAS側から調整

    (例)delay = 150

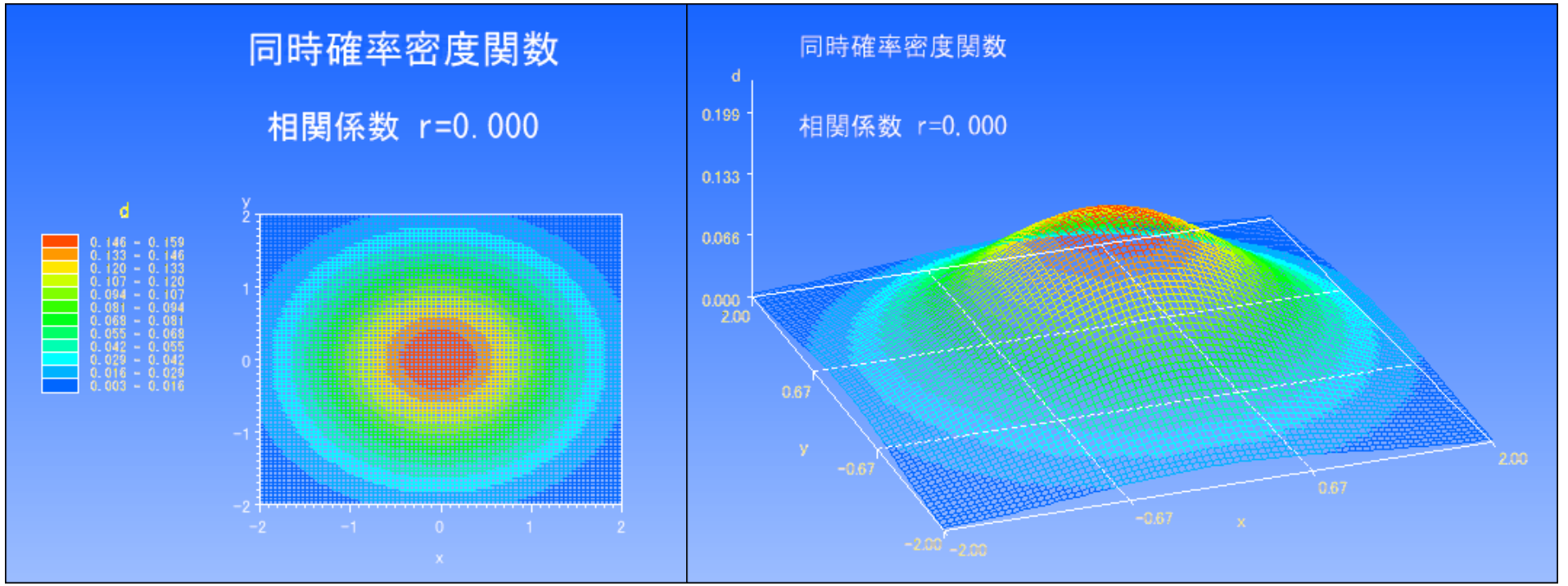


図2. 二次元正規分布(高品位)

GIFのメリット

- パワーポイントに貼り付けられる
  - スライドショーで駆動
  - プレゼンテーションの最中に動画を見せることが可能
- 複数のグラフを同時に駆動
  - スライド内で同期は取れる
  - 異なるタイプのグラフの連動を見ることができる

動的な分割表

- カイ2乗検定
  - 各セルの期待値からの乖離を色で表現
  - 期待値からの乖離は標準化残差とする

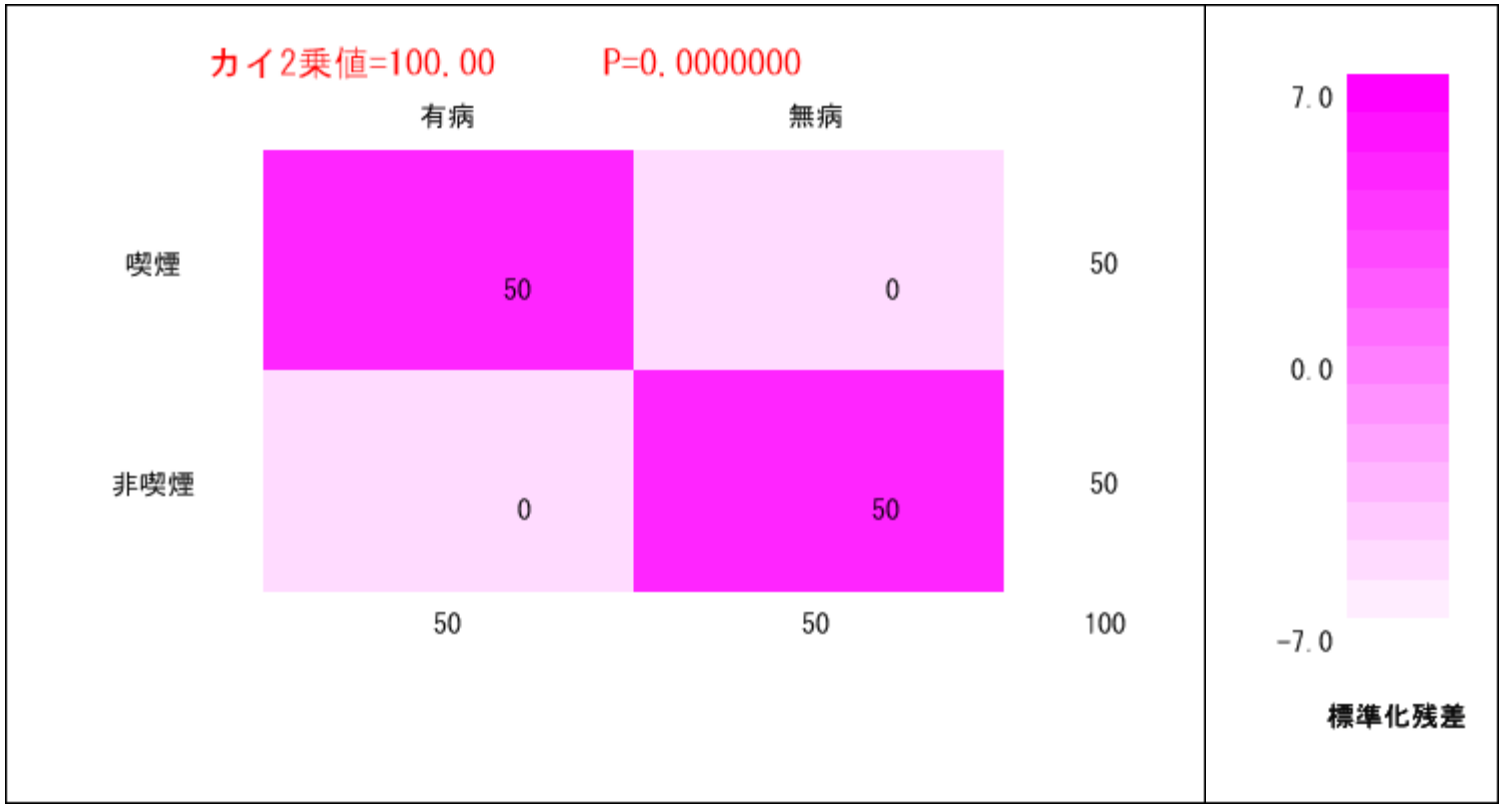

カイ2乗値=100.00
P=0.0000000
有病
無病
喫煙
非喫煙
50
0
0
50
50
50
50
100
7.0
0.0
-7.0
標準化残差

図3. カイ2乗検定

動的な分割表（2）

- 層別カイ2乗検定
  - 2表（日本人・日本人以外）を同時に動かす
  - 2表の傾向性はBreslow-Day検定にて監視
  - シンプソンのパラドクスを表現

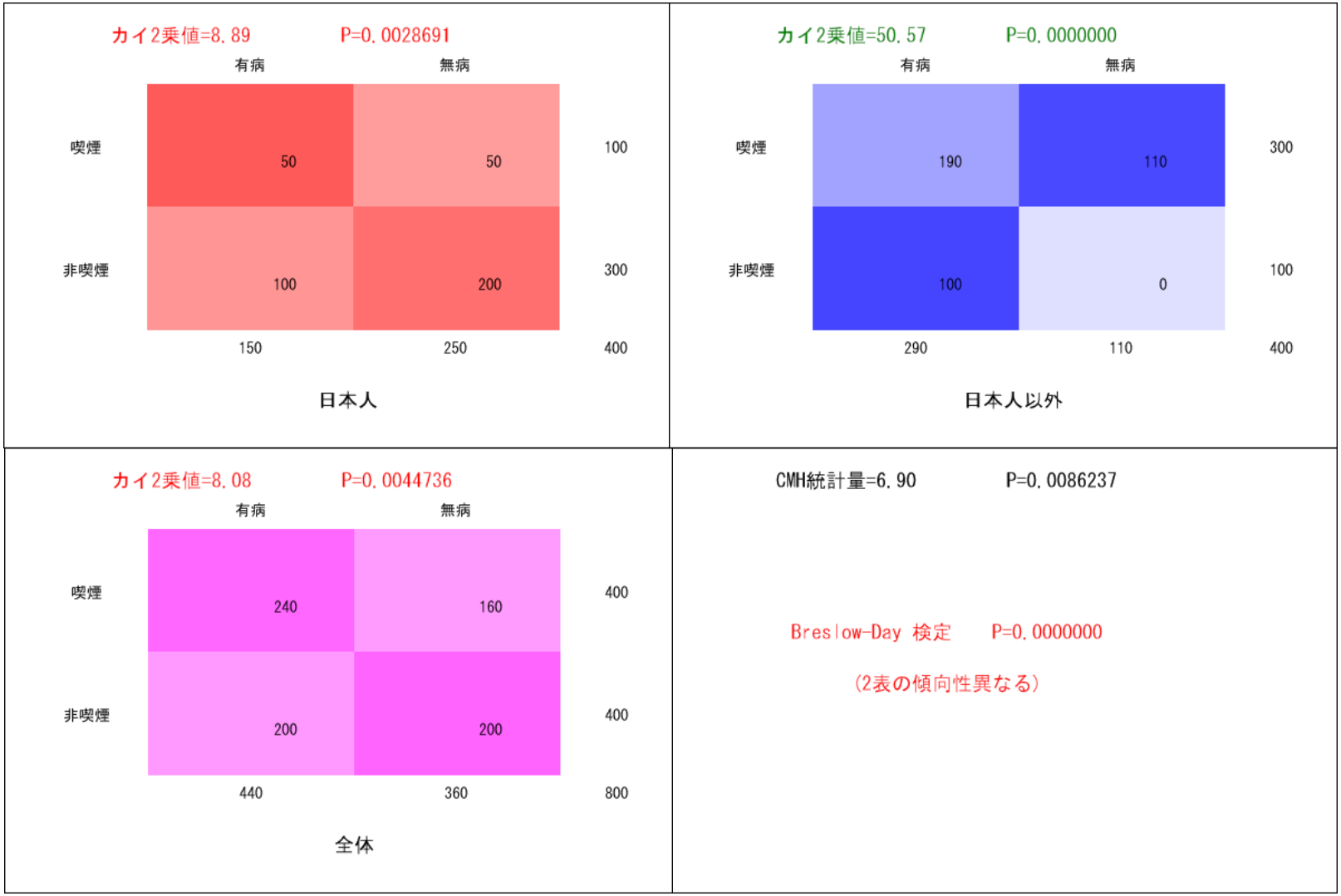
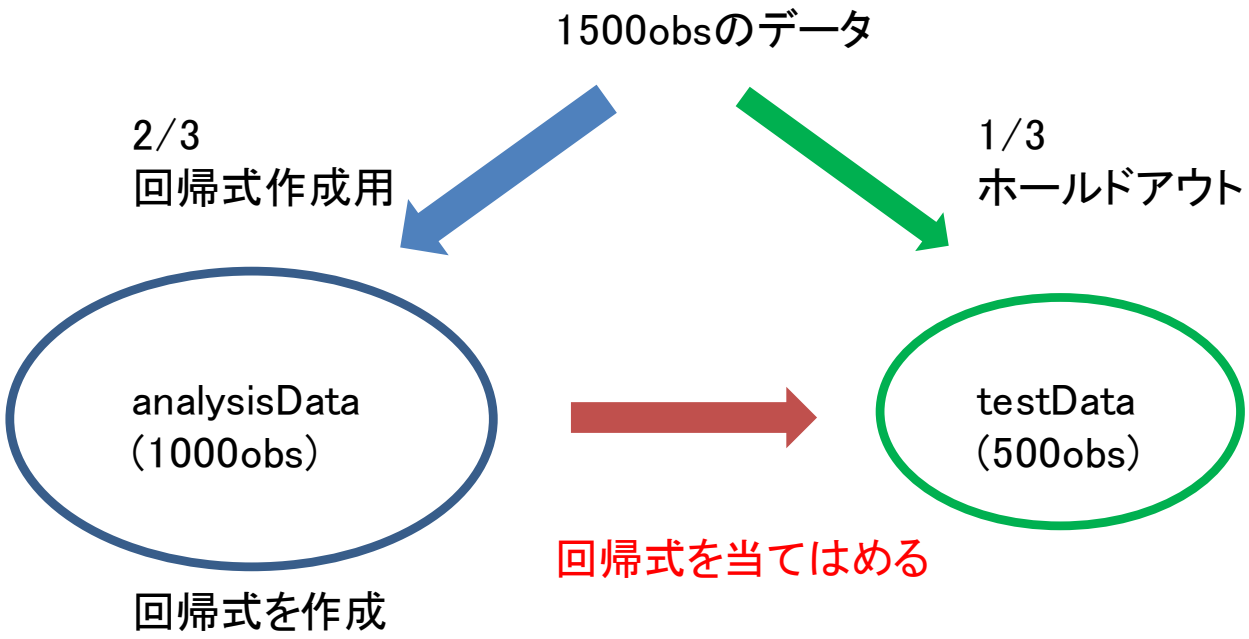
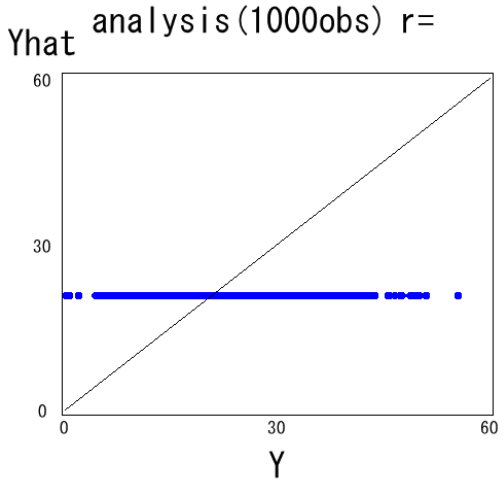

カイ2乗値=8.89 P=0.0028691
有病 無病
喫煙 50 50 100
非喫煙 100 200 300
150 250 400
日本人
カイ2乗値=50.57 P=0.0000000
有病 無病
喫煙 190 110 300
非喫煙 100 0 100
290 110 400
日本人以外
カイ2乗値=8.08 P=0.0044736
有病 無病
喫煙 240 160 400
非喫煙 200 200 400
440 360 800
全体
CMH統計量=6.90 P=0.0086237
Breslow-Day 検定 P=0.0000000
(2表の傾向性異なる)

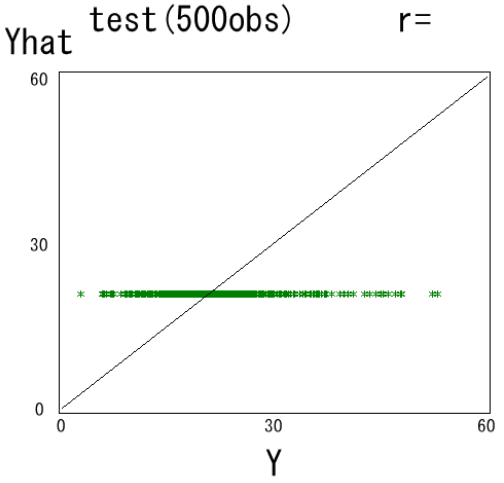
図4. 層別 カイ2乗検定

動的な折れ線

- ASE（残差平方和をN数で割ったもの）を比較







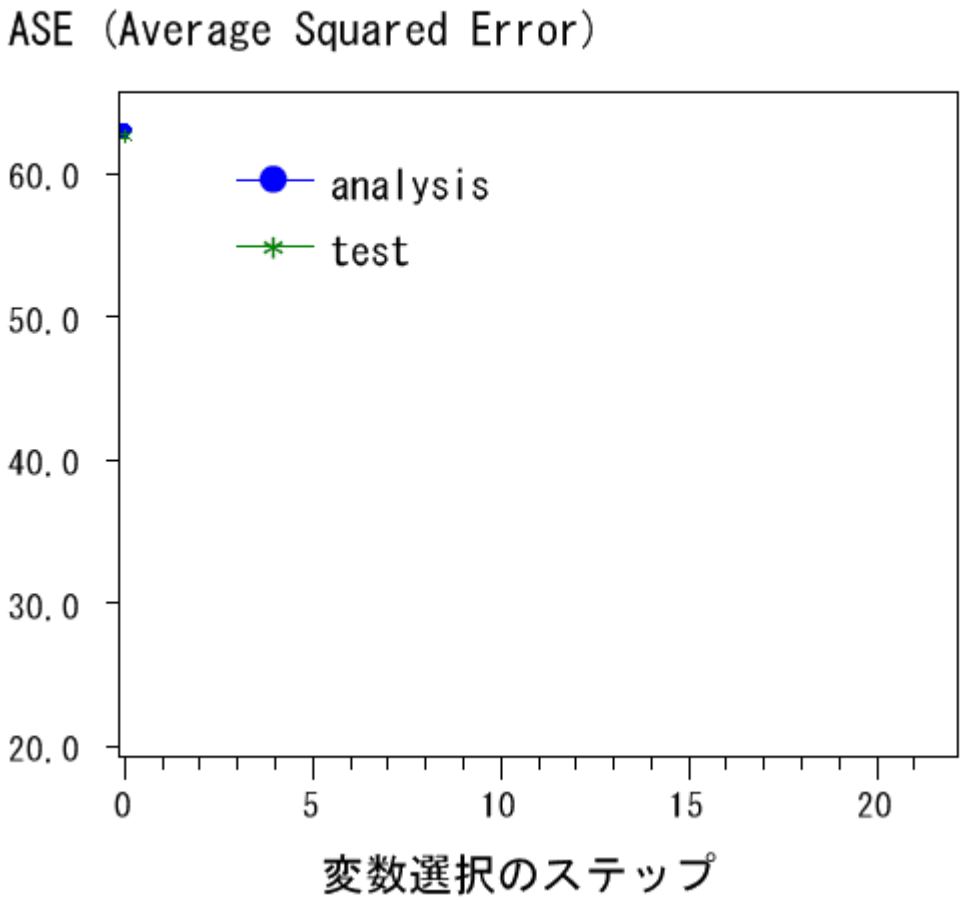
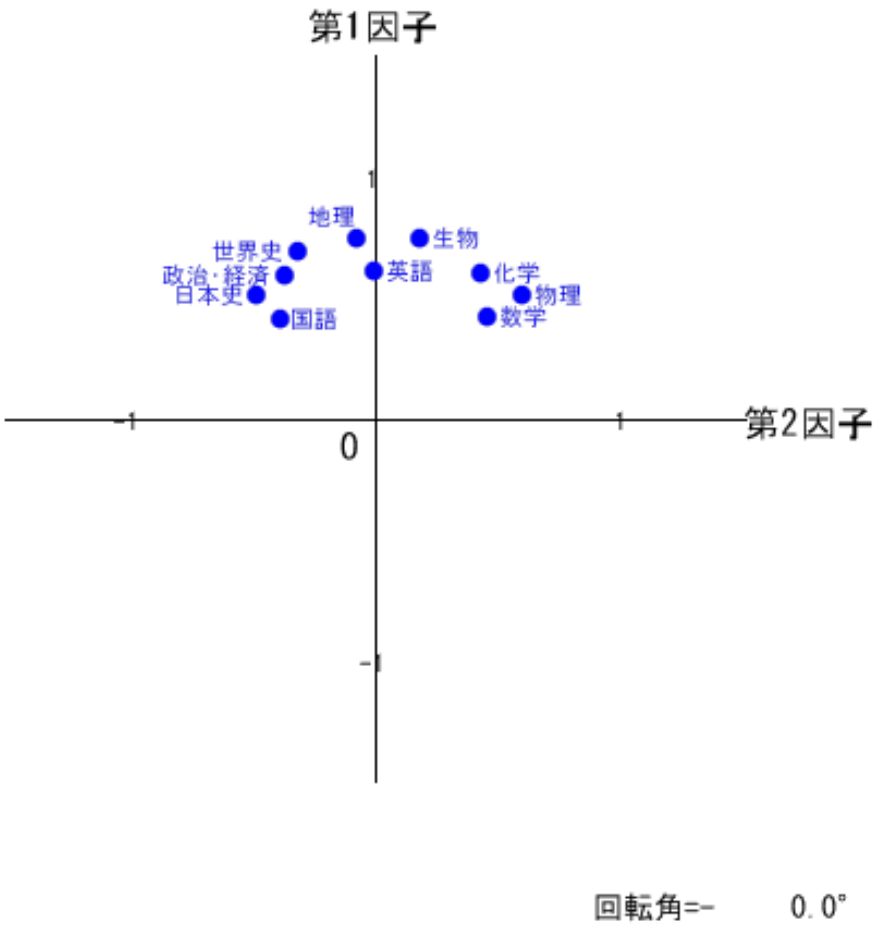


図5. 線形重回帰分析における過学習の概念

動的な座標軸

- 因子軸の回転（バリマックス基準）
  - 使用データ（10教科50人分の人工データ）

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



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

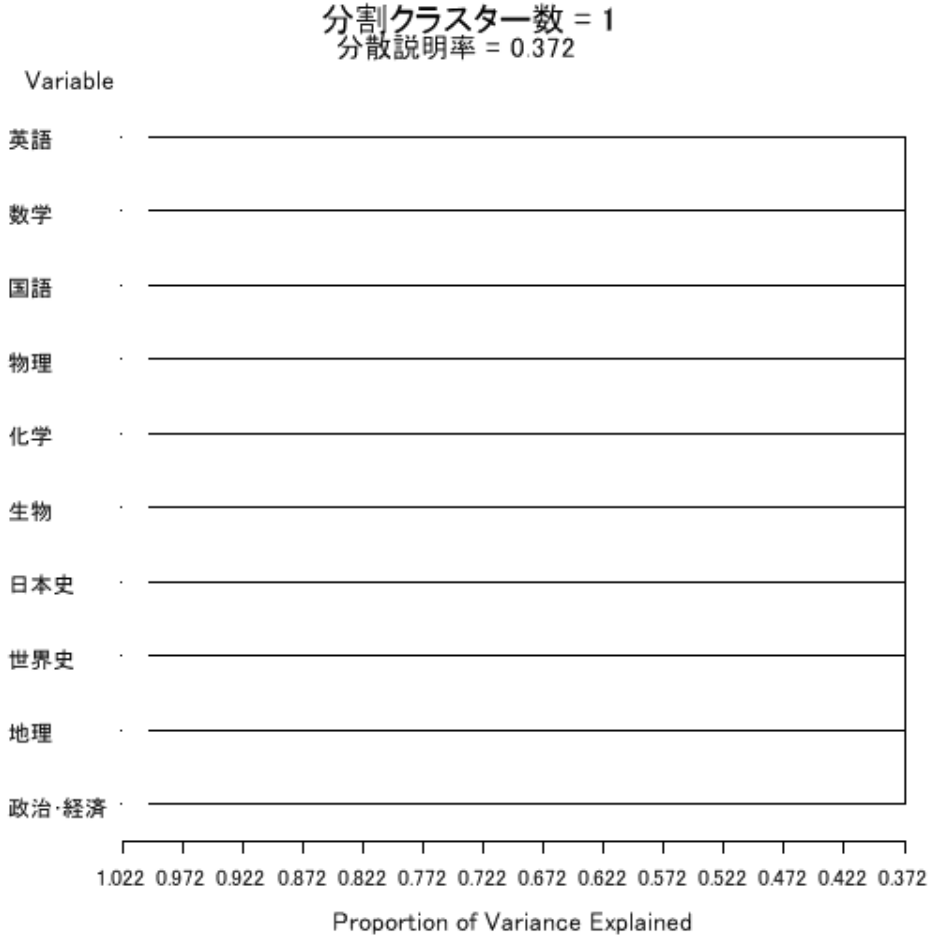
回転後因子負荷量

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

動的なデンドログラム

- VARCLUSによる変数のクラスタリング
  - 分散説明率と固有値との関係を判り易く表現
  - 使用データは図6に同じ



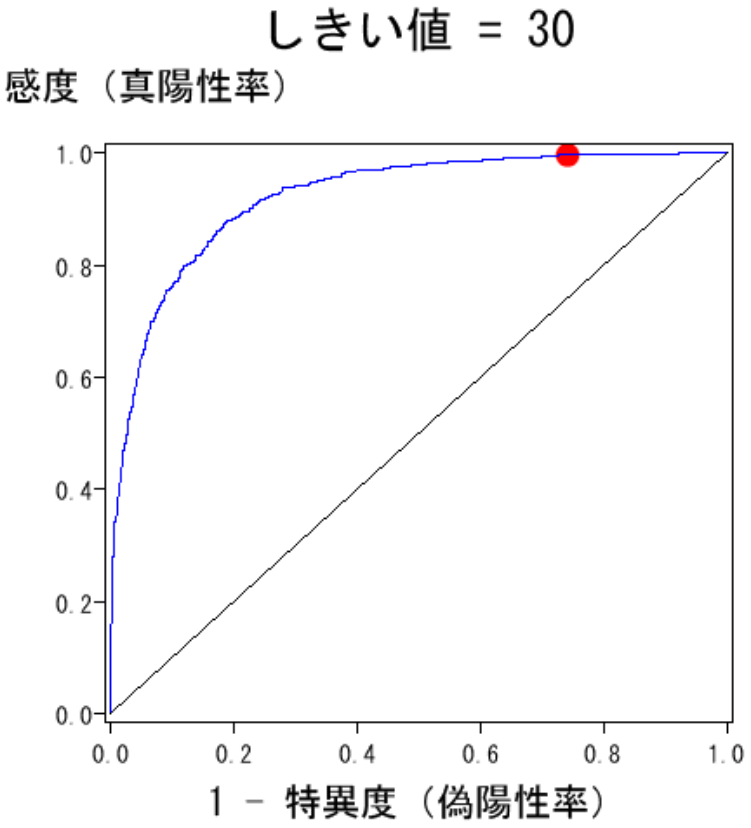
主成分分析の第1固有値

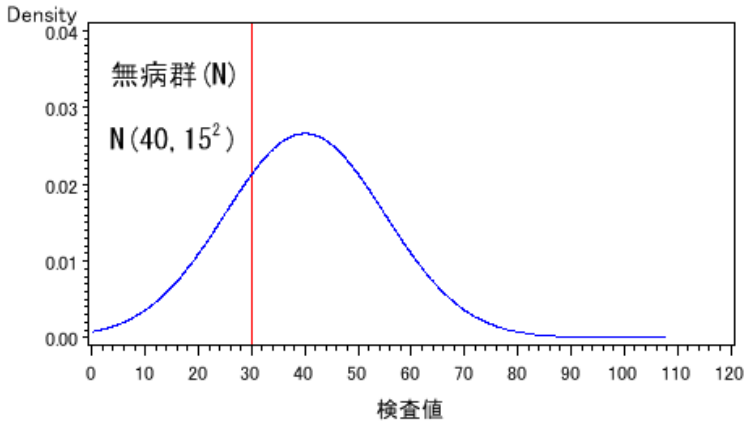
第1クラスター	英・数・国・物・化・生・日・世・地・政	3.722

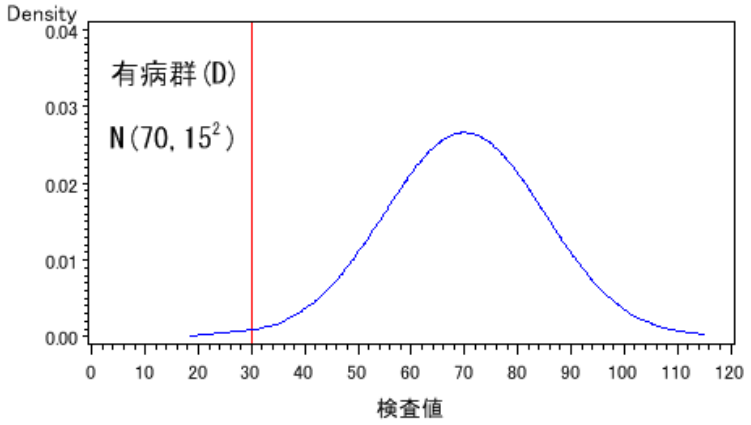
図7. 変数の階層的クラスター分析

動的なしきい

- ある臨床検査薬を考える
  - 有病群は正規分布$N(70,15^2)$に従う(500例)
  - 無病群は正規分布$N(40,15^2)$に従う(9500例)
  - しきい値を30から上昇させ、しきい値を超えた場合を陽性(+)とみなす







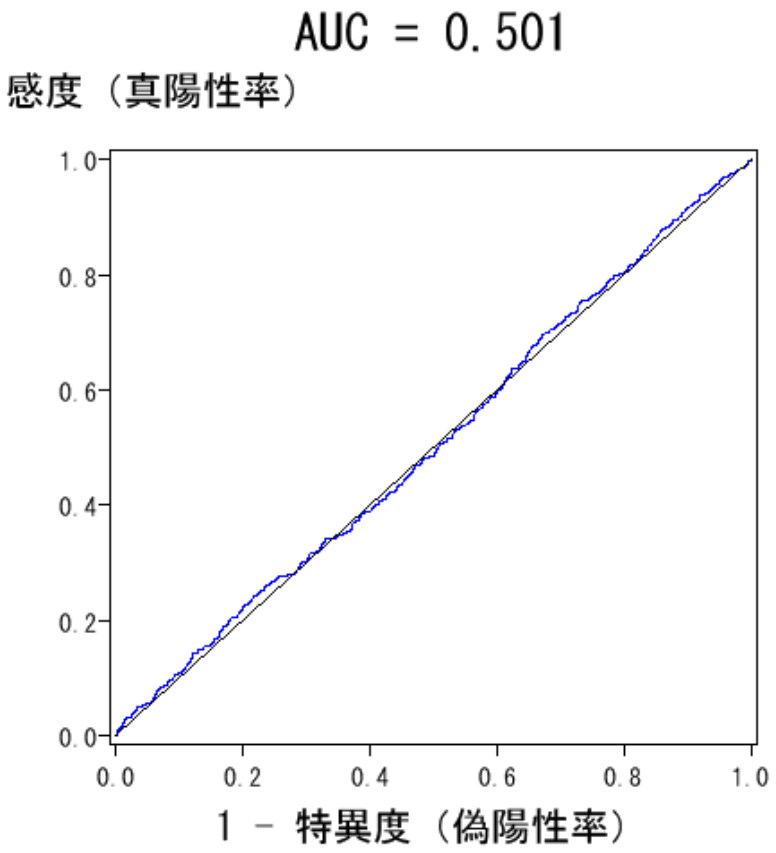
P(−|N) = 0.257

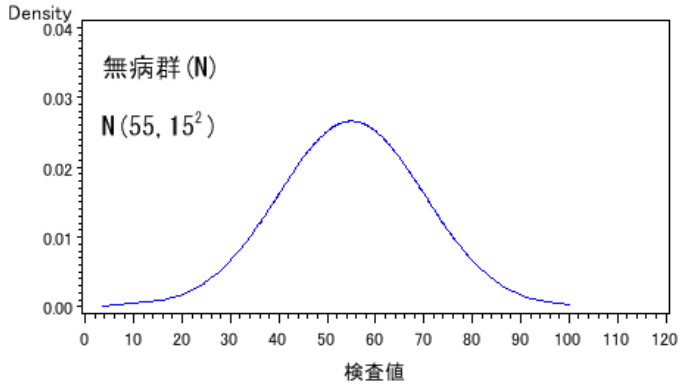
P(+|D) = 0.998

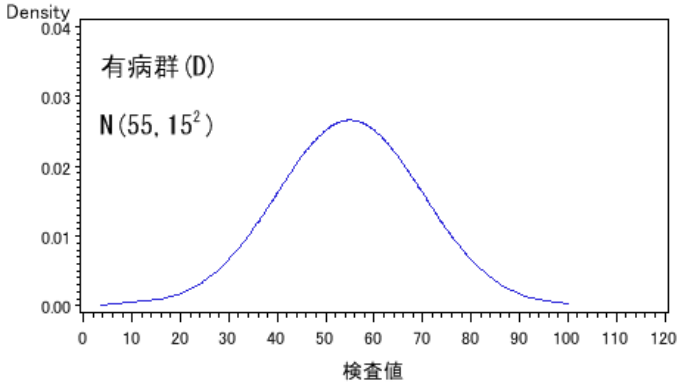
図8. ROC曲線としきい値

動的なROC

- 図8と同様のデータにて
  - 有病群は分散を変えずに平均を55→80に上昇
  - 無病群は分散を変えずに平均を55→30に下降
  - ROC曲線のAUCを計算し、群間差との関係を見る







群間差の t 統計量 = 0.76

図9. AUCと群間差の関係

SGグラフの動画化の手順

```
ods graphics / reset imagename="WRK" ;
proc lifetest data=LIFE ;  ～

data ANNO;
  ～
  imgpath="WRK.png"; style='fit'; output;
run;

proc ganno anno=ANNO; run;
```

外部ファイルWRKに
吐き出す

外部ファイルWRKを
ANNOTATEデータセット化

ANNOTATEデータセットを
Gグラフ内に呼び込む

動的なカプランマイヤー図

- ログランク検定とウィルコクソン検定の比較
  - プラセボ群と実薬群で生存率が逆転するデータでのカプランマイヤー図の比較

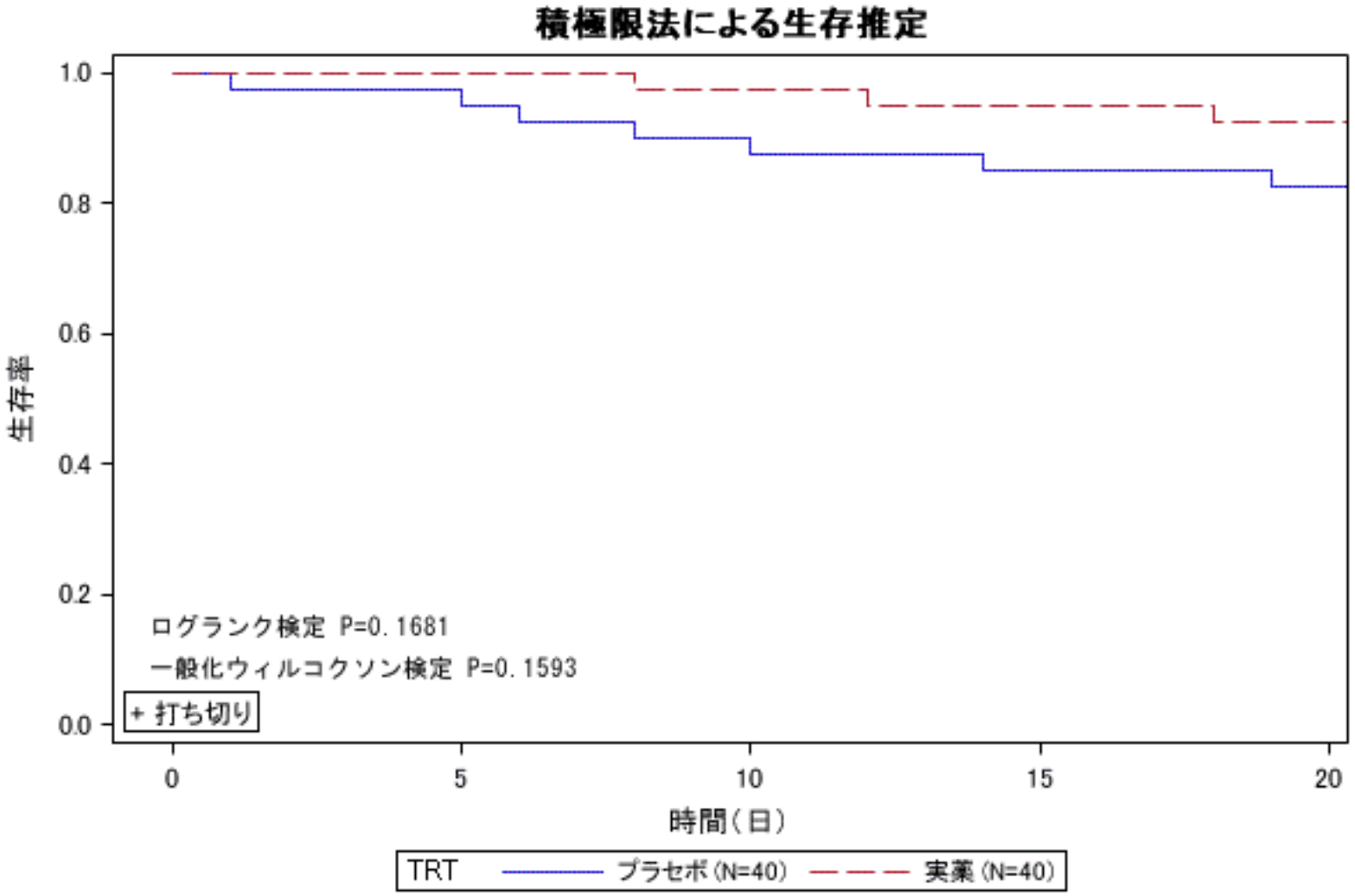
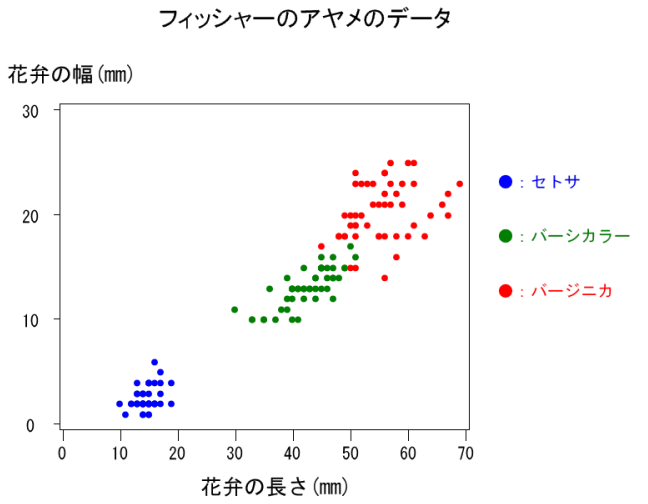


図10. ログランク検定とウィルコクソン検定の比較

動的なクラスター

- FASTCLUSによる3分割
  - フィッシャーのアヤメ（花弁の幅・長さ）を使用



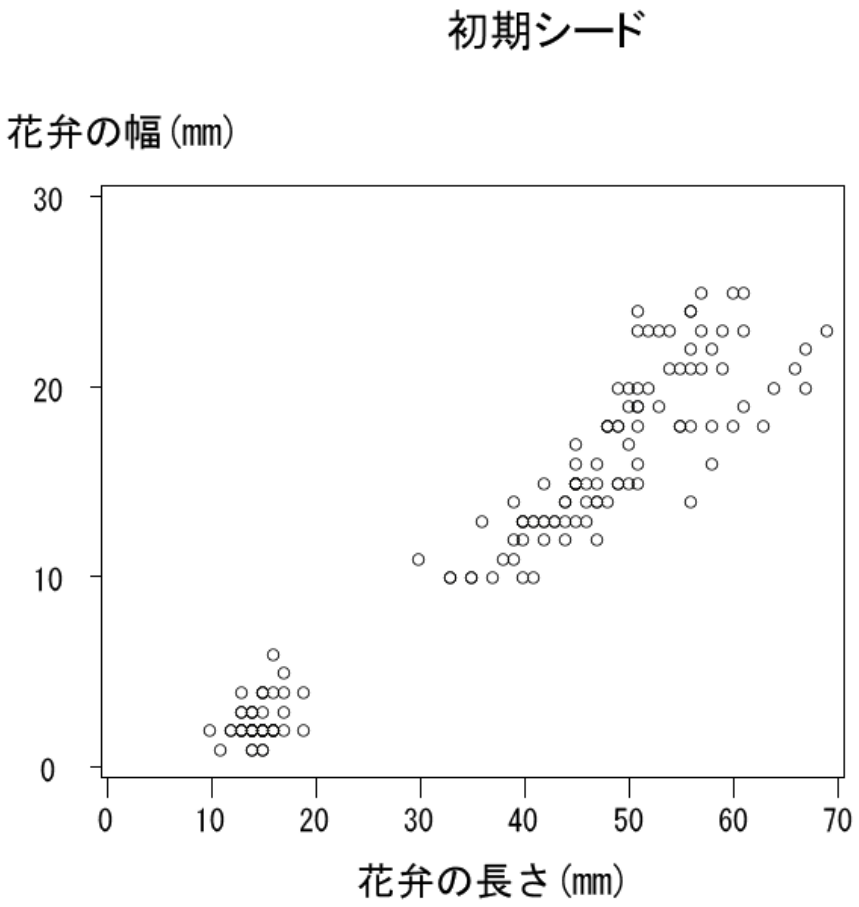


図11. k-meansクラスタリングの概念

まとめ

- 動画は大量の情報を1枚に集約できる
- 動画は結果だけでなくプロセスを説明可能
- 動画は統計量の多面的な連動を表現する

最後に