統計教育と統計ソフトの共生

新村秀一
成蹊大学 経済学部

Collaboration of statistical education and statistical package

Shuichi Shinmura
Dept. of Economics, Seikei Univ.

要旨

- SUGI に,「15年ぶり?」の復帰.
- 正規の統計教育を受けずに,統計ソフトを教師として,実践的な統計の知識を獲得した過程を紹介.
- 信念:問題解決学として生きた知識獲得の王道
  - 「専門家が必要とする機能を全て備え,初心者から専門家までが使いやすい理数系のソフトがあれば,それを習得し実際の問題解決を心掛ける」ことこそ,『幾何学の王道』.
  - 統計ソフトは,「統計学の最良の個人家庭教師」になる.
- 定年を1年後にひかえ,自分史をまとめた.
- 2010年に完成した自分の一生の研究である「最適線形判別関数」の,次の発表のイントロ.

キーワード: 統計学の理解,統計ソフトの役割,統計学の学習法,SASとJMPの重要な機能

1. SASとJMPの書籍を例に

1.1 一生の研究テーマとの出会い（STN春号参照）

- 心電図の自動診断解析システム
  - 1971年に大学を卒業し，SCSからNECに出向し，大阪府立成人病センターで「心電図の自動解析システム」のプロジェクトに参加.
  - 野村裕医師から32個の心電図の異常所見と正常所見の数千件のデータを渡され，「計量診断学（高橋晄正編）」を勉強し，診断論理の研究を行う.
    - 米国IBMの研究者が，心電図の波高値から判別分析で診断論理を研究
    - 心電図，電気工学，統計の勉強
- 枝分かれ論理に対する敗北
  - 4年間研究しても，野村医師の開発した枝分かれ論理（決定木分析，Expert System)にかなわなかった.

敗北の総括

- 地球モデル

[25]新村秀一(1984). 医療データ解析，モデル主義，そしてOR，OR誌，29-7，415-421.
[26]新村秀一(1985). 医学における診断とは－枝分かれ法からAIへ－，OR誌，，30-8，501-507.

  - 医療の正常/異常は，Fisherの考える2群判別でない
  - 正常所見は地球であり，異常群はある計測値が連続的に大きく（あるいは小さく）なり，地球から噴き出た山脈.
    - 異常の典型例は平均でなく山頂，判別境界付近に多くの症例
    - MT理論（マハラノビス田口）
      - 異常は正常（基底空間）からのマハラノビスの距離の大きなもの
    - 1クラス判別分析：SVMの研究者は，MT理論を知らないのか？
- スペクトル診断 （[27] 新村秀一，北川護，高木義人，野村裕(1973). 二段階重みづけによるスペクタル診断，第12回日本ME学会大会論文集，107-108.）
  - Bayesの定理で，「地球モデルの新手法」を提案
  - しかし，ロジスティック回帰がそれを簡単に可能

1.2 書籍文化とソフト文化の違い

- 統計書を独学する問題点
  - 行間が読めない
    - Seal, H. L. (1964) Multivariate Statistical Analysis for Biologists . Methuen & Co., Ltd. , [塩谷実訳(1970). 多変量解析入門ー生物学を題材にしてー. 共立出版.]
    - 配置行列
      - 後年, SASのマニュアルで配置行列の生成法を知り氷解
    - 重回帰分析, 分散共分散分析, 数量化Ⅰ類などが統一的に理解できる
  - どこまでいっても, 完結しない
    - 世界最高水準のSASの出力を, 体系的に理解し, レポートできればそれで良いとの割りきり・・・精神衛生に良い
    - 統計手法を体系的に理解し, 優先度をつける

1.3 SASによる回帰分析の実践[1]

- J.Sall副社長のテクニカル・レポートの翻訳
  - MATRIX言語で, 行列表現による統計アルゴリズムの理解の重要性(NLINの手法の紹介)
  - 最小二乗法からLpノルム回帰まで
  - 多重共線性
    - [28] 新村秀一(1988). データ解析に見るグラフ. OR誌, 33-4, 172-178.
- 付録に掃き出し演算子の紹介
  - Goodnight社長のテクニカル・レポートの翻訳
  - 変数選択法が, 行列の掃き出し演算子で解説
    - [29]新村秀一(1983). 重回帰分析における掃き出し演算子. OR誌, 28-11, 565-569.
- 統計量の意味が, 統計書と異なり, 具体的で明確

1.4 統計手法の体系化

表1 データを調べる

	連続尺度	名義尺度，順序尺度
1変数	基本統計量とヒストグラム	度数表（クロス集計）
2変数	散布図，相関，単回帰	分割表（クロス集計）
3変数以上	クラスター分析，主成分分析	多重クロス集計

表2 予測手法

		目的変数	
		連続尺度	名義尺度/順序尺度
説明変数	連続尺度	重回帰	判別分析 ロジスティック回帰
	名義尺度 順序尺度	分散分析	FUNCAT

1.5 SASの解説書

- SAS言語入門[6]：SAS言語の文法書
- 統計処理エッセンシャル[4]，（JMP[10]）
  - 「学生の成績データ」による入門書
  - 40人，8変数（SN,成績，勉強時間，支出，飲酒日数，喫煙の有無，性別，クラブ活動）
  - 最適線形判別関数の研究に利用
- 易しく実践 データ解析の進め方[5]，（JMP[11]）

[40]新村秀一(1986). 科学万博データの解析，オペレーションズ・リサーチ，30-12，754-766.

  - 科学万博の184日間の時系列データ
  - 変数（入場者数，ごみの排出量，迷子数等，17変数）
  - ほぼ重要な手法を解説
- 上級編は出版できず

1.6 JMPの解説書

- **ウイーンの思い出（2003/4～2004/3)**
  - **「JMPを用いた統計およびデータ分析入門[9]」の監修**
  - **「JMP活用 統計学とっておき勉強法[10]」**
    - 手計算を前提に，4件*2変数のデータで統計量の紹介
    - 学生の成績データをJMPで分析
    - JMPの評価版の添付
    - 革新的なサイトライセンス契約
- **JMPによる統計レポート作成法[11]**
  - **科学万博**
  - **統計レポートの作成に重点**
- **上級編が出版できず**

1.7 パソコンによるデータ解析

- 新村秀一（1995）．講談社ブルーバックス．
- シカゴのファースト・イリノイ銀行が，従業員から性差別，人種差別をしていると訴えられた裁判のデータ．
- SPSSのサンプル・データ（BANK.SAV）
- 見どころ
  - 単純な分析では，銀行側が敗訴
  - 職種別に層別し，現在の給与を初任給で回帰すれば差別はない
  - 年齢を上手く順序尺度にすれば，分割表の結果が良い．

2. 統計上の研究と貢献

2.1 大阪府立成人病センターとの研究
2.2 日本医科大学
2.3 介護保険と決定木分析
2.4 個人による統計ソフトの開発者との論争
2.5 某銀行の投資分析システム

2.1 大阪成人病センターとの研究

- 「Fisherの線形判別関数を超えて[8]」の発表
- 心筋梗塞の予後予測
- 疫学調査部は，がん検診などの整備されたデータがあり，自由に統計研究に利用させてもらえた.
  - 胃がんと乳がんに関する各種判別分析の成績評価をROC曲線で行った[12－14].
  - ROCを医学診断に応用したのはL.B.Lusted医師で，彼の解説書を最初の研究指導者の野村医師[15]が翻訳.
  - 筆者は，それを個別の判別関数の判別境界を変えた評価と，異なった判別関数の評価に用いることを提案.
    - JMPでは，ロジスティック回帰の評価にROCが利用
    - LDFやQDFにも利用すべき?・・・間違った研究の予防

2.2 日本医科大学

- 鈴村産婦人科教授(CPDデータ:児頭骨盤不均衡)
  - 帝王切開/自然分娩の手術法の決定に,X線画像から「鈴村氏法」と呼ばれる簡易診断法を開発.
  - 19個の計測値で判別したいという研究テーマ.
  - RSQUAREプロセジャーで2^{19}(約52万個)のモデルを計算.
    - 3個の多重共線性があり
    - 16変数をフルモデルとして考える [16－18].
- 丸山ワクチン(SSM)の分析
  - 約30万症例から,手術を受けた患者さんに限定し,術後3ケ月以内,6ケ月以内,9ケ月以内,1年以内の4層に分け,生存日数を検討したところ,術後3ケ月以内に投与開始した患者群が,有意に1年以内の患者群の生存日数より長い[19].
    - 医療情報学会で発表.丸山ワクチンの認可に影響を及ぼすかもしれないということで,医事関連の業界紙のインタビュー.
    - しかし発表前日,現状維持の決定がなされ,騒ぎは収まった.

2.3 介護保険と決定木分析

- 大学に移る2年前(1994),土肥医師(高校の1年先輩で,当時東京都の医療部門No.2)から,介護保険システムを開発するために集めた「1分間タイムスターディ」の分析法の相談を受けた.
  - 主成分回帰とか色々と試みたが上手くいかない.
  - そこで,「患者の医療行為に当てられた分数を目的変数にし,患者属性を説明変数として回帰木を行い,得られた葉ノードは介護時間でセグメント化されているので,その枝分かれ論理をC言語でシステム化」することをアドバイス.
  - 私より情報処理能力に優れた彼は,CHAIDを用いて分析し,得られた枝別れ論理をあっという間にC言語に落とした.
  - あまりに結果が良いので,統計分析の歴史的な社会システムへの貢献と考え,大学に移る前年に大会長を務めた日本計算機統計学会で発表[20].
  - 発表後,「国会で承認されれば在宅まで含めて実施」を聞かされた.
    - 特養などの病院データなので,在宅まで拡張するのは母集団が異なり問題では?」
  - 大学に移った年,彼がサンデープロジェクトで「このシステムは僕が作った問題だらけのシステムです」と暴露.
    - 在宅での痴呆老人の問題に加え,
    - CHAIDで順序尺度を名義尺度で分析したため「順位の逆転」.

決定木分析の注意点

- 野村医師が「枝分かれをそのまま使っても上手くいかない．確信度とともに，非確信度も計算し，ある値以上になれば，上の階層の別の分岐に入れている」という言葉.
- その後，息子が卒業研究を自由に選んで良いと指導教授に言われたのでテーマをほしいという.
  - 「CHAIDと2分岐のCARTの比較」
  - 「回帰木は一元配置の分散分析で，分類木は分割表で停止則を考える」というテーマを与え
  - OR学会で，卒業記念がわりに発表させた[21].
    - 多分岐は，必ずしも2分岐より良いわけではない
    - 息子の勉強を見たのはそれが最初で最後です.

2.4 個人による統計ソフトの開発者との論争

- 丹後さんとの統計ソフトに関する質問論争[22]
  - SASデータセットを超えた3次元データが操作できる
  - SASで実現可能
- SALSの開発者と非線形回帰に関する論争[23]
  - SASのNLINの収束域
  - SAS/MATRIXで記述し，NLINと合わせ分析
- SASの普及に尽力していて，九州大学の研究グループからつけられたあだ名が「SAS坊や」.
  - 鈴木教授(筑波大)から，「少しでもSASに批判的なことをいうと私が出てきて，大学の研究者仲間で評判が悪い」.

丹後さんとの論争

- 丹後他(1980). 医療データ解析のための統計パッケージSPMSの開発. 医用電子と生体工学. 18/2. 30-35.
  - SPMSの3次元構造を扱える優位性を，BMDP，SPSS，SASと比較.
  - 住商コンピュータサービス：SASの紹介(新村） (1979)を引用.
    - ユーザーがDATAステップでプログラミングや，ユーザー作成プロセジャーが登録できると言っているが，マニュアルに記述されていない.
- 新村(1980). 丹後論文に対する質問. 医用電子と生体工学，18/6，60-62.
  - DATAステップのプログラミング機能，複数DATAセットをマージして3次元あるいは変数に時系列情報で実現化. PROCによる繰り返し処理. MATRIXによるユーザー作成，DATA stepのプログラミング機能，Supplementalの紹介.
- 丹後，刈谷(1980). 新村氏の質問に対する回答. 医用電子と生体工学. 18/6. 63-64.
  - 質問に名を借りたSASの紹介である.

中川・小柳「非線形最小二乗法のソフトウエア」

- 中川・小柳(1982). 非線形最小二乗法のソフトウエア. 情報処理，23/5，442-450.
  - 著者らの開発したSALSと，SASなどと比較.
  - Marquardt法が概して安定で計算回数も少ない.
  - SASの最急降下法やDUD法は問題がある
    - これはSASの問題でなく手法の問題である.
- 新村(1984). 中川・小柳「非線形最小二乗法のソフトウエア」についての討論－SASの評価について－. 情報処理，25/7，697-703.
  - SASのNLINとMATRIXでガウス法，マルカート法，DUD法を，収束域の限界まで広げて検証し，中川らの表と比較.
  - マルカート法＞ガウス法＞DUD，DUDは開発途上で悪い.
- 中川・小柳(1984). 非線形最小二乗法ソフトウエアの収束のテストについて－新村氏のコメントに対する回答. 情報処理，25/7，703-707.
  - 新村の検証は，自分たちの結論を追認したものである.

2.5 某銀行の投資分析システム

- 某銀行がSAS/IMLで投資分析システムを開発
  - 実データで検証すると，組み入れ比率が0などに収束
- 筆者が，LINDO，Speakesyと検証すると，係数の最大値/最小値が10^8以上の場合，必ずエラー.
  - SAS/IMLは，研究用には適しているが，システム開発言語でない.
  - 数理計画法は，SASデータセットのようなフラットテーブルには適していない
- 数理計画法LINDOのテキストに明記
  - 証券会社のシステムでも経験
  - 他の開発でも経験

3. 統計レポートの作成(成蹊大学にて)

これまでの成蹊大学の統計教育の中で，情報科学Ⅱ（2年次配当，統計ソフトを使った実習）を中心に報告する.

3.1 SASによる教育（1996年～2001年）

3.2 SPSS（～2006年ごろ）

ウイーン：2003/4～2004/3

3.3 JMP(2007-2011)

3.1 SASによる教育（1996年～2001年）

- 2年次配当の情報科学Ⅱを前期・後期に開講
  - 定員は，PC室の制限で当初は60人．この時代情報教育の黎明期で，他に競合ソフトによる開講科目がなかったので，ほぼ定員を満たす．
  - 定員制限のため受講に漏れて次年度に受講する者もいた．
  - テキストは[4]を用いた．「学生の成績データ」を用いて，SASの操作法と出力結果の解釈を小説のように解説．
  - 評価は，授業で習ったことを参考に自分でデータを決めて20頁以上の統計レポート
- 授業のトラブル
  - 多くの学生が自宅でPCを利用していないので，入力速度に格差
  - SAS言語[6]のコマンド入力間違いが頻発
  - 課題提出に際しては，この時代インターネット上に分析に耐えるデータがないので，書籍のデータを手入力
  - 試験前にレポート作成が集中し，PC室の利用がネック
  - 優秀な学生からは，「他の授業の倍以上の学習時間を必要とするが，終われば達成感で満足しています」などというコメント．

（1）高杉君：「日本・死者急増」
　データを以下のように各種調査票から収集し，最終的に次の回帰式を得た．
zikosuu＝14.531054+0.012590＊menkyo＋e
****
データ収集の出典の一例を次のようで多岐にわたっている：
・運輸省自動車交通局技術安全部管理課編 平成6 年9 月度車種別自動車保有車両数月報
・財団法人自動車検査登録協力会編 平成6年3 月末現在市区町村別自動車保有車両数，輸入車保有車両数
・（社）全国軽自動車協会連合会編市区町村軽自動車車両数による．
・都道府県別保有台数は，乗用車（普通車，小型車，軽四輪車），トラック（普通車，小型四・三輪車，被けん引車，軽四・三輪車），バス（普通車，小型車），特殊用途車両（消防車，救急車，タンク自動車などの普通車，小型四・三輪車，大型特殊車）及び二輪車で構成される．この数値には駐留軍人，軍属の私有車，外国人の私有車などを含むが，防衛庁関係の車両は含まれていない．
****

（2）田中君：「プロ野球選手の成績」についての分析
　彼は，私の授業が厳しいという風聞のため，4年生になって単位を全て取った上で受講．
余裕があるのか，30頁の統計レポートとPower Pointの発表資料も提出．
　また，参考文献に[5]を挙げているので，この時期はテキストをグレードアップしたようだ．
　そして，次の重回帰式を導いた．
年俸＝-5405.51＋3888.93＊リーグ＋77.11＊打点＋165.64＊四球
　その後，大学院に進学し，新聞社の懸賞論文の特選を受賞している．

3.2 SPSS(～2006年ごろ)

- SASの利用者が私一人ということで，SASの契約がキャンセルされ，SPSSを利用
- SASのCUIから，SPSSのGUIに代わって学生の負担が軽減
- それ以上に，総理府統計局などのHP上に質の高いExcelデータが開示，学生の負担は少なくなった.
  - この時期，筆者のHP(http://sun.econ.seikei.ac.jp/~shinmura/)に，次の2名のレポートを載せている.
  - どのようなレポートでS評価になるかの参考に供した.
- HPの更新は浅井(GLP担当)夫人に頼んでいたが，研究費に余裕がなく久しく更新していない.

(3)芦川さん:「日本映画産業についての分析」
- 授業ではめだたない学生が，25頁の秀逸なレポート提出
- その後，時系列データは経済環境による年代区分が明確であることが分かった

(4)前田さん:「日本の住宅事情に関するレポート」
- 入学後のテストに合格し，2年次配当の「情報科学Ⅱ」にとび級で入ってきた1年生.
  - 2-3回目の授業後，顔面蒼白な顔で「先生の話す日本語の意味が全く分かりませんので，ドロップアウトしたい」といってきた.
  - 聞くと1年生なので，もう少し辛抱すれば理解できるようになるし，授業後分からないことを質問してきなさいと言って返した.
- まったく統計の知識がないにもかかわらず，25頁のレポートを提出

(5)高橋君:「プロ野球選手の推定年俸についての統計分析」
- 筆者の3年のゼミ生
- 過去の秀逸なレポートの中から田中君と同じテーマを希望.
  - データを自分で集め直すこと，
  - 最低2年以上のデータで重回帰式の予測と検定を行うことの条件で認めた.
  - 最終的に35頁のレポートを提出してきた.
- 食品会社に就職し，深夜にExcelで1000品目以上の販売データの分析をしていて助けて下さいという.「なぜJMPを使わないの」と聞くと，「会社が買ってくれません」
  - 2か月の週次の8個の販売データから今月の計画をExcelで立てる
  - 「JMPの評価版は20件まで入力し処理できるので，なぜそれで分析し，会社にJMPのメリットをアピールし，商用版を買ってもらう努力をしないの?」といって打ち切った.
  - 生活の知恵や，気点を利かすことまで教育できないもどかしさを感じた.

3.3 JMP(2007-2011)

- JMPのユニークな点は，ライセンスと操作法の革新性
  - 従来の統計ソフトを授業に用いる場合，必要な統計手法と統計量のオプションをきめ細かく教える必要．
  - JMPのメニューは，手法がわずか7個のカテゴリー(プラットフォーム)にまとめられている．
  - 分析する変数を選び，[Y列]をクリックし[OK]すれば，グラフから統計量すべてが出力される．
- 授業では重要なものだけを説明し，不用な出力は無視することを宣言し，統計ソフトの操作法にさく時間が節約
- 学生にとって，統計ソフトの操作法や分析データが容易に手に入るようになったが，以前のように秀逸なレポートがない
  - ゆとり世代にとって，負担の大きな授業を避ける傾向と，
  - 他のPCを使った授業が増えてきたためである．
- その中で，秀逸なものを紹介する．

(6) 坂本君:「2007年日本の優良企業分析」

- 坂本君は平均給与の高い優良企業に就職したい
- 日経優良企業100社と自分でインターネットで調べて非優良企業30社を選定し，財務データなど23変数を集めた．
- 彼のレポートの2章に見るように3つの作業仮説を検証した．

****

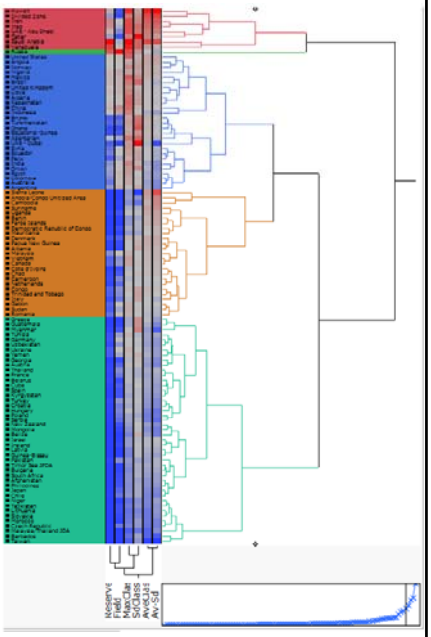
2 レポートの目的と作業仮説

レポートの目的は，優良企業の特徴を優良企業でない企業と比較することによって明らかにすることである．作業仮説は，①優良企業はそうでない企業よりも財務的に優れている．②従業員一人当たりの利益が大きい企業は平均給与が高い．③利益が高い企業は株価が高い，の3つを検証する．

****

(7)井上さん:

- 社会人大学院生
- 初めて統計ソフトを利用
- 修士1年:金融データの分析
- 修士2年:「油田規模分布の統計分析」



4. 統計入門教育

4.1 統計入門の概略

2010年:授業15回

2011年:3.11のため授業が15回から11回

4.2 試験の結果

2011年の方が良い

4.3 散布図の検討

問題学生の検討

4.1 統計入門の概略

- テキストは[10]を用いた
- 本書は，第1部では(x,y)=(0,1),(1,1),(1,3),(2,3)という2変数*4件の簡単なデータで，表の統計量を説明.
- 第2部では「学生の成績データ」を用いて，JMPの出力結果を用いて統計量の意味を説明.

週	2010年	2011年
1	PowerPointで概論	同左
2	最頻値，中央値，平均値	同左
3	範囲，四分位範囲,SD,CV	同左
4	学生データの解釈	同左
5	正規分布	同左
6	自由度，SE，ｔ分布	相関係数
7	中間試験	中間試験
8	相関係数	9回目
9	Excelで相関の計算	10回目
10	単回帰分析	12回目
11	単回帰分析	期末試験
12	分割表と独立性の検定	
13	分割表と独立性の検定	
14	補講	
15	期末試験	

4.2 試験の結果

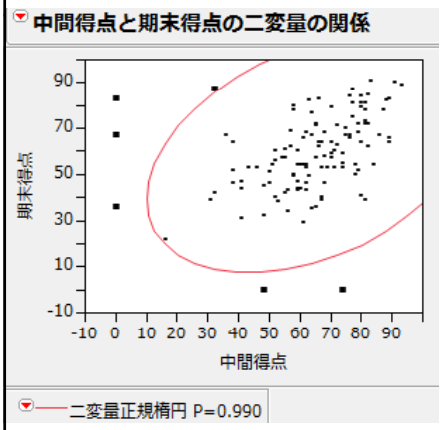
- 2010年度の中間試験は，　48点(10%点)で合否判定したが，わずか6問(6点)で判定可.
- 2011年の期末試験は，最高点は8点も上昇し，半舷授業を克服し好成績.
- 2011年度の期末は，中間に比べ、最高点は11点も高い．中間試験が80点で，期末試験を99点を取った学生を呼び出し勉強法を確認したところ，半舷授業の悪影響を克服して好成績な理由は次のように考える.

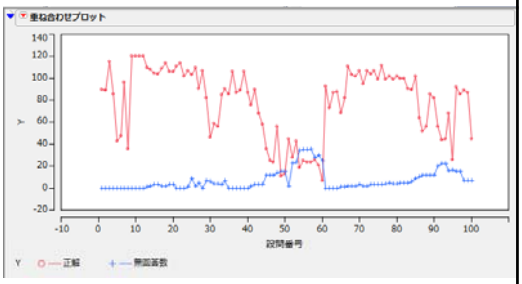
1) 講義ノート等の資料を事前に開示した工夫,
2) 多くの学生がテキストやPowerPointによる事前配布資料の予習を行ったこと,
3) Excelの計算式の開示が良かったと考える.

統計家は，あらゆるデータを分析すべし.

		2010年度		2011年度	
		点	次元	点	次元
中間	10%点	48	6	42	12
	50%点	66	12	61	15
	90%点	82	13	79	9
	最高点	93		88	
期末	10%点	40	12	43	8
	50%点	60	12	60	13
	90%点	82	11	81	8
	最高点	91		99	

4.3 散布図の検討





5. 研究に用いたデータ

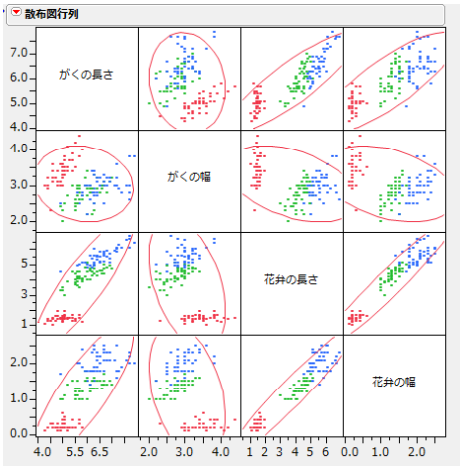
4種類の実データ(教師データ)

- Fisherのアイリスデータ(Fisherの前提を満たす)
  - 100ケース，4個の説明変数からなる2種類 (versicolor, virginica).
  - 15 個のモデルがある.
- CPD (Cephalo Pelvic Disproportion) データ(多重共線性)
  - 自然分娩群180症例，帝王切開:60症例
  - 52 万個のモデル($=2^{19}-1$)のうち，40(あるいは26)個で検討.
- Flury & Rieduel (1988)が集めたスイス銀行紙幣データ(MNM=0)
  - 真札と偽札各100枚.
  - 6個の計測値があり，63 個のモデルで評価.
  - 2変数(X4，X6) でMNM=0.
- 学生 データ(一般位置にない)
  - 35 人の合格と15人の不合格.
  - 5個の 説明変数.
  - 31個のモデル.

評価データ: 4種の実データから，正規乱数(Type1)と復元抽出(Type2 , Type3)でリサンプリング標本を作成

全ての研究は，149個(=15+63+40+31) あるいは135個のモデルで検討.

5.1 Fisherのアイリスデータ



- セトサ，バーシクル，バージニカの各50件
- 4個の計測値
- セトサは他の2群から線形分離可能
  - 統計研究家は，NM=0の問題は，判別が容易と誤解する原因
- Fisherの仮説を比較的満たす
- 統計入門の良い教材（相関係数の危険性）

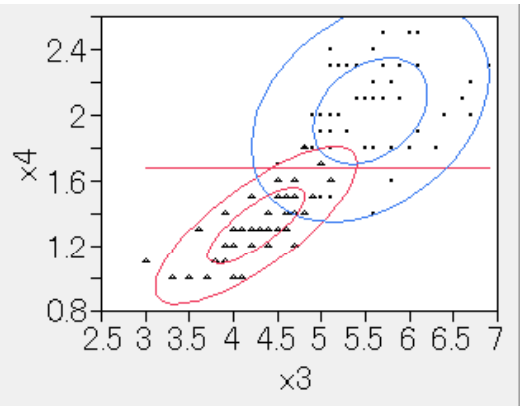
アイリスデータ

- 4変数の全モデル（15個）で判別
- F5は事前確率が等しい，LDFのNM.
- Q5は2次判別関数のNM. ＊は分散共分散が等しいという仮説の適合度検定.
  - 赤色のモデル
    - QDFが薦められているが，
    - LDFの誤分類数が少ない.
  - これは，他の分析例にも多数
- 各誤分類数をMNM で単回帰

MNM =0+MNM ,
*LP=1.329+0.912*MNM ,*
*F5=1.471+1.014*MNM ,*
*Q5=1.477+1.080*MNM*

p	AIC	IP	F5	Q5	説明変数
1	-250	5(6)	6	6*	X4
1	-231	5(7)	8	7	X3
1	-163	24(27)	27	30	X1
1	-145	29(37)	42	42	X2
2	-261	3(5)	5	7*	X2 X4
2	-260	3	6	3**	X3 X4
2	-252	3(4)	6	6	X1 X3
2	-248	5	6	5**	X1 X4
2	-233	5(6)	7	10	X2 X3
2	-161	24(25)	25	29	X1 X2
3	-276	2	4	4**	X2 X3 X4
3	-273	2	3	3**	X1 X3 X4
3	-261	3(4)	5	6*	X1 X2 X4
3	-251	2	7	8	X1 X2 X3
4	-282	1	3	3**	X1 X2 X3 X4

Model 6の散布図



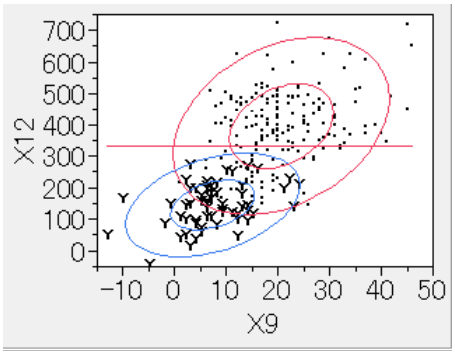
100重交差検証法によるLDFの平均誤分類確率は3.25%だけ，ロジスティック回帰のそれは2.19 %だけ，改定IPLP-OLDFの平均誤分類確率より悪い

5.2 CPDデータ

X1	年齢	X11	入口部横径
X2	経産回数	X12	X13−X14
X3	仙骨の数	X13	入口部面積
X4	入口部前後径	X14	児頭面積
X5	かつ部前後径	X15	子宮底
X6	狭部前後径	X16	腹位
X7	最短前後径	X17	外結合線
X8	児大横径	X18	大転子間径
X9	X7−X8	X19	側結合線
X10	入口部前後経		

- 児頭骨盤不均衡の胎児をもつ180人の自然分娩群と，60人の妊婦
  - 分娩前に計測値から手術法を判別
  - 鈴村氏法の妥当性を判別分析で検証
- X9とX12は，日本医科大学産婦人科教授の開発した「鈴村氏法」を検証するため，他の計測値の差
- 3個の多重共線性がある
- RSQUAREを利用

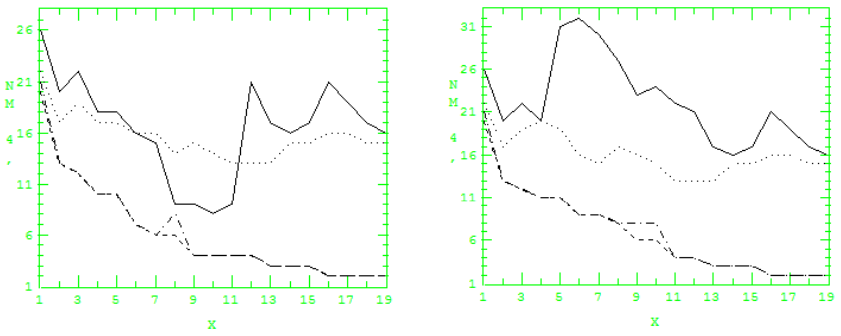
CPD



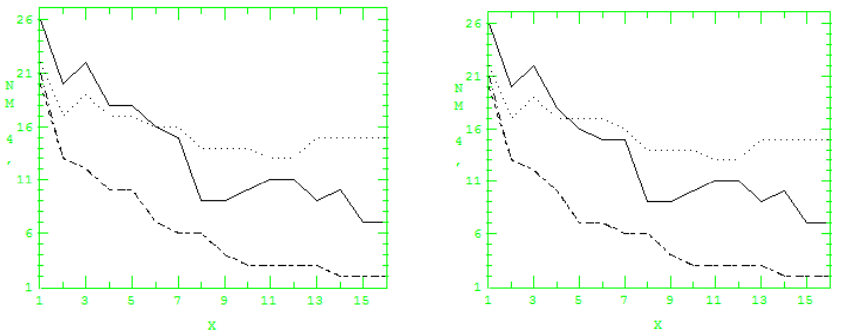
- (X9,X12)の散布図
- Yは帝王切開群，・は自然分娩群：アンナカレーニナの例えに反する
- 100重交差検証法で，LDFの平均誤分類確率は7.26%,ロジスティック回帰は2.13%だけ，改定IPLP-OLDFより悪い．

P	Rank	Type	IP	FP	QP	AIC	P	Rank	Type	IP	FP	QP	AIC
1	1	FBfb	19(20)	23	22	-568	9	3	fb	4	14	9	-587
2	2	FBfb	13	17	20	-586	10	1	B	6	15	24	-586
3	3	FBfb	12	19	22	-587	10	6	F	4	14	8	-586
4	1	Ffb	10	17	18	-589	10	*	fb	3	14	10	-585
4	3	B	11	20	20	-587	11	1	B	4	13	22	-586
5	1	Ff	10	17	18	-589	11	*	F	4	13	9	-587
5	2	b	8(7)	17	16	-588	11	*	fb	3	13	11	-584
5	3	B	11	19	31	-588	12	1	FB	4	13	21	-584
6	1	B	9	16	32	-589	12	*	fb	3	13	11	-582
6	2	b	7	17	15	-588	13	1	FB	3	13	17	-582
6	3	Ff	8(7)	16	16	-588	13	*	fb	3	15	9	-580
6	*	DOC1	13(12)	20	20	-587	14	1	FB	3	15	16	-581
6	*	DOC2	11	19	22	-587	14	*	fb	2	15	10	-578
7	1	B	9	15	30	-589	15	1	FB	3	15	17	-579
7	2	Ffb	7(6)	16	15	-588	15	*	fb	2	15	7	-576
8	1	F	6	14	9	-588	16	1	FB	3(2)	16	21	-577
8	2	B	8	17	27	-588	16	*	fb	2	15	7	-574
8	5	fb	6	14	9	-587	17	1	FB	2	16	19	-575
9	1	B	6	16	23	-587	18	1	FB	2	15	17	-573
9	2	F	4	15	9	-587	19	1	FB	2	15	16	-571

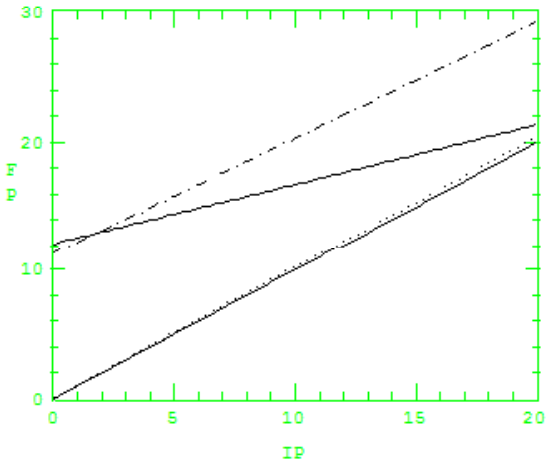
CPDの19変数の誤分類数の比較
左：上昇基本系列上の誤分類数（実線は2次判別関数，破線はLDF，
大きな破線は改定IP−OLDF,一点鎖線は改定LP−OLDF）
右：下降基本系列



CPDの16変数
左：上昇基本系列上の誤分類数（実線はQDF，破線はLDF，
大きな破線は改定IP-OLDF,一点鎖線は改定LP-OLDF）
右：下降基本系列



IPの定義域で,「***改定IP-OLDF＜改定LP-OLDF ＜LDF＜2次判別関数***」の順に誤分類数が増える

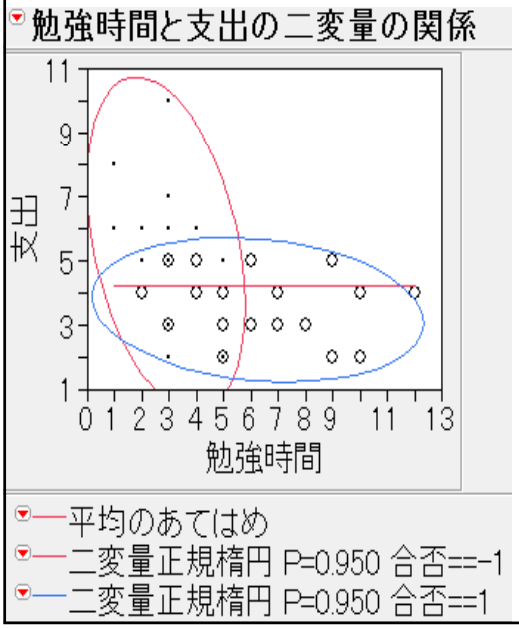


5.3 学生の成績データ

- 学生40人の成績を，5変数で判別
  - SAS, SPSS，JMP, Statisticaの教科書に用いる
  - 変数選択法は2変数を選ぶ

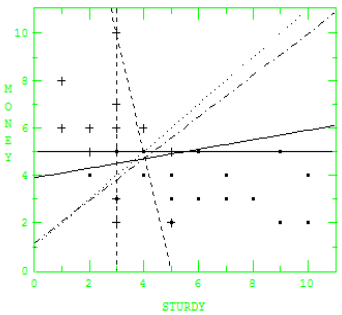
項	推定値	SE	t値	p値	VIF
切片	0.394	0.603	0.654	0.518	.
勉強時間	0.131	0.059	2.232	0.032	1.774
支出	−0.096	0.100	−0.957	0.345	2.322
飲酒日数	−0.169	0.117	−1.446	0.157	3.078
性別	−0.072	0.255	−0.282	0.780	1.205
喫煙の有無	−0.029	0.279	−0.103	0.919	1.450

学生の散布図



- 100重交差検証法で，LDFの平均誤分類確率は7.09%，ロジスティック回帰は5.37%，改定IPLP-OLDFより悪い

一般位置にないデータ



- 一般位置にないデータ
  - Harr条件を満たさない
  - 数理計画法では退化
  - 配置行列の小行列が退化
  - 判別超平面上に(p+1)個以上のケースがくる
- 水平な実線：改定IP-OLDF（$M=10^6$）
  - 新版のIP-OLDFとLP-OLDFはほぼ重なる.
- 右上がりの実線：改定IP-OLDF（M=30）
- 一点鎖線： LDF
- 破線： 改定LP-OLDF
- 大きな破線の垂線：旧版のIP-OLDF
  - 判別超平面上に10人の学生
- 右下がりの破線：LP-OLDF
- 散布図（注：＋は不合格，□は合格）

6個の判別関数

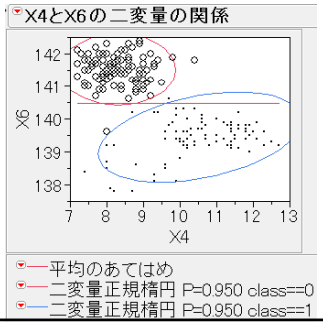
	勉強時間(x)	支出(y)	定数項	Y=ax+b
改定IP-OLDF (M=10^6)	2	-3.3E-5	1.66E6	Y=6E-6*x+4.999967
改定IP-OLDF (M=30)	2	-10	39	Y=0.2*x+3.9
改定LP-OLDF (M=1)	0.5	-0.5	0.5	Y=x+1
IP-OLDF(旧版)	0.333333	0	-1	X=3
LP-OLDF(旧版)	0.2	0.04	-1	Y=-5*x+25
LDF	-0.66004	0.74987	-0.885	Y=0.8802*x+1.1800

5.4 銀行紙幣データ(MNM=0)

	変数名	記号	内容	説 明
1	length	X1	横幅長	紙幣の横の長さ
2	left	X2	左縦幅長	紙幣の縦の長さ(左側)
3	right	X3	右縦幅長	紙幣の縦の長さ(右側)
4	bottom	X4	下枠内長	紙幣の下端から内側の枠までの長さ
5	top	X5	上枠内長	紙幣の上端から内側の枠までの長さ
6	diagonal	X6	対角長	対角線の長さ
7	class	Y	真偽	札の真偽 (1: 真札, −1:偽札)

- Fluryら(1988)が集め
- 判別分析の本で利用
  - しかし，(X4,X6)でMNM =0を認識していない．
- MNM の単調減少性 $MNM_p \geqq MNM_{(p+1)}$
- (X4,X6)を含むすべてのモデルでMNM =0
- アイリスとCPDデータの平均値間の距離を拡大し，MNM =0に変換データでも確認

項	推定値	標準誤差	t値	p値	VIF
切片	24.090	6.551	3.677	0.000	.
length	0.017	0.030	0.563	0.574	1.286
left	-0.117	0.044	-2.689	0.008	2.517
right	0.110	0.040	2.771	0.006	2.618
bottom	0.150	0.010	14.769	0.000	2.176
top	0.157	0.017	9.222	0.000	1.906
diagonal	-0.209	0.015	-13.901	0.000	3.033

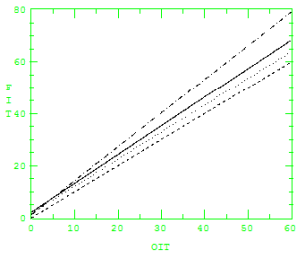


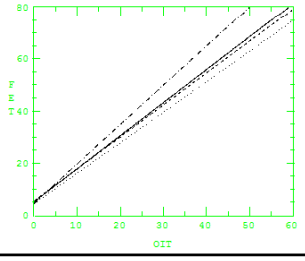
- 6変数モデルで検討しても問題が見当たらない
- X4とX6の2変数の散布図
  - 100重交差検証法で，LDFと改定IPLP-OLDFの平均誤分類確率の差は0.61%，ロジスティックの差は0%.
- なぜ線形分離が認識できなかったのか？

変数選択法で選ばれた5変数モデル

変数	AIC (上昇)	AIC (下降)	推定値	SE	t値	p値	標準β	VIF
切片	-275.3		26.26	5.29	4.96	0.00	0.00	
diagonal	-603.9		-0.21	0.01	-14. 0	0.00	-0.48	2.94
bottom	-699.5		0.15	0.01	14.96	0.00	0.43	2.10
top	-775.6		0.16	0.02	9.22	0.00	0.25	1.90
right	-776.0		0.11	0.04	2.87	0.00	0.09	2.58
left	-781.1	-781.1	-0.11	0.04	-2. 64	0.01	-0.08	2.34
(length)		-779.4						

5.5 2変量正規乱数による評価データの検証





Speakeasyで2変量正規乱数

*X=NORMRANDOM(ARRAY(400, 1:))*2;*

Y=NORMRANDOM (ARRAY(400, 1:));

A群の平均を原点に固定し, 0度, 30度, 45度, 60度, 90度で回転.

B群のXの値には0から8までの整数i, Yに0, 2, 4の整数jを加えて平行移動.

5*9*3個の組合せの教師と評価データを作成.

上は教師:一点鎖線(LP-OLD), 実線(LDF), 破線(QD), 破線(MNM)

「**IP-OLDF＜QDF＜LDF＜ LP-OLDF**」

下は評価データ

QDF＜IP-OLDF＜LDF＜ LP-OLDF」

6. 終わりに

- 28歳の時, SASを統計の先生とし, SASを使って統計の受託計算を開始.
  - SASを習得する過程で, 従来の統計書にない新しい視点で出版.
  - SASのミニコン版の販売を通し, 製薬と金融機関に販売し, 統計ユーザーの拡大.
  - 大学に移り自分の統計習得の過程を授業で再現
  - JMPのサイトライセンス契約と評価版がユーザー拡大に貢献.
- 来年に定年を迎え自伝的な内容になったが, SAS普及に関連した多くの仲間は正確を期すため本人が自伝を語ることが後世の役に立つと考える.
- 統計に続いて数理計画法や数学も同じようにソフトウエアを教師とし勉強.
  - 統計と数理計画法を融合した最適線形判別関数を考えることで, 「Fisherの線形判別関数の問題点」を克服できた.
  - 大学に移ることで, 定年前に自分の一生の研究を完成できた.
- 経営科学の学部や学科は理系と文系にあるのに, 数理計画法を含むソフトの普及が統計スフとに比べて遅れている.
- 1980年代の統計ソフトの熱気は?
- 自分の弟子と呼べる人を育てることができなかったことが大きな反省点.

文献

[1]J.P.Sall (新村訳)(1986). SASによる回帰分析の実践. 朝倉書店.

[2]J.H.Goodnight(1976).Computation Methods in General Linear Models. Proceedings of the Statistical Computing Section,ASA,Washington,D.C.

[3]森村英典・牧野都治編(1984). 統計・OR活用辞典. 東京書籍.

[4]高森寛・新村秀一(1987). 統計処理エッセンシャル. 丸善.

[5] 新村秀一(1989). 易しく実践 データ解析の進め方. 共立出版.

[6]新村秀一(1994). SAS言語入門. 丸善.

[7] 新村秀一(1993). 意思決定支援システムの鍵ー有り余るコンピュータ・パワーをどう使うー. 講談社.

[8] 新村秀一(2012).Fisherの判別分析を超えて. 2012年SASユーザー会論文集.

[9]J.Sall他(2004). JMPを用いた統計およびデータ分析入門. SASジャパン.

[10]新村秀一(2004). JMP活用統計学とっておき勉強法. 講談社.

[11]新村秀一(2007). JMPによる統計レポート作成法. 丸善.

[12]新村秀一・鈴木隆一郎・中西克己(1981). 胃X線像の各種判別分析. オペレーションズ・リサーチ, 26-1, 51ー60.

[13]S.Shinmura, T. Suzuki, H. Koyama & K. Nakanishi (1983). Standardization of medical data analysis using various discriminant methods on a theme of breast diseases. MEDINFO 83, J.H. Van Bemmel, M.J, Ball and O. Wigertz editors, 349-352, North-Holland Publishing Company.

[14]新村秀一・鈴木隆一郎・中西克己(1983). 各種判別手法を用いた医療データ解析の標準化 ― マンモグラフィによる乳癌の診断 ―. 医療情報学, 3-2, 38-50.

[15] L.B.ラステッド(野村裕/中村正彦 訳)(1976). 臨床診断への新しい道 意思決定の理論と実際.コロナ社.

[16]新村秀一, 三宅章彦(1983). 重回帰分析と判別解析のモデル決定(1)ー 19変数をもつC.P.Dデータの多重共線性の解消 ―. 医療情報学, 3-3, 107-124.

文献

[17]新村秀一(1996). 重回帰分析と判別分析のモデル決定(2)ー19変数を持つC.P.D.データのモデル決定ー. 成蹊大学経済学部論集, 第27巻第1号, 180-203.

[18]新村秀一(2004). 数理計画法を用いた最適線形判別関数(8)ー524,287個の回帰モデルの検討ー. 成蹊大学経済学部論集34-2,53-70.

[19]新村秀一・飯田和美・丸山千里(1987). SSM(人型結核菌鯛抽出物質, 丸山ワクチン)の癌治療における帰無仮説モデルによる評価. 医療情報学, 7-3, 263-276.

[20]土肥徳秀・筒井孝子(1995). 提供されたケアからみた要介護高齢者のタイプ分け. 第9回日本計算機統計学会シンポジューム, 161-170.

[21]新村秀樹・新村秀一(2002). 決定木分析のモデル選択に関する考察(1). オペレーションズ・リサーチ春季研究発表会, 142-143.

[22]新村秀一(1980). 丹後論文に対する質問. 医用電子と生体工学, 18-6, 455-456.

[23]新村秀一(1984). 中川, 小柳「非線形最小二乗法のソフトウェア」についての討論 ― SASの評価について. 情報処理, 25-7, 697-707.

[24] 新村秀一(1999). パソコン楽々数学. 講談社.

[25]新村秀一(1984). 医療データ解析, モデル主義, そしてOR, オペレーションズ・リサーチ, 29-7, 415-421.

[26]新村秀一(1985). 医学における診断とはー枝分かれ法からAIへー, オペレーションズ・リサーチ, 30ー8, 501-507.

[27] 新村秀一, 北川護, 高木義人, 野村裕(1973). 二段階重みづけによるスペクタル診断, 第12回日本ME学会大会論文集, 107-108.

[28] 新村秀一(1988). データ解析に見るグラフ. OR誌, 33-4, 172-178.

[29]新村秀一(1983). 重回帰分析における掃き出し演算子. OR誌, 28-11, 565-569.

[30]新村秀一(1986). 科学万博データの解析, オペレーションズ・リサーチ, 30-12, 754-766.