データとソフトウェアを活用したデータ分析教育で何をどう教えるか？
－予測型分析を例に－

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
椿　広計

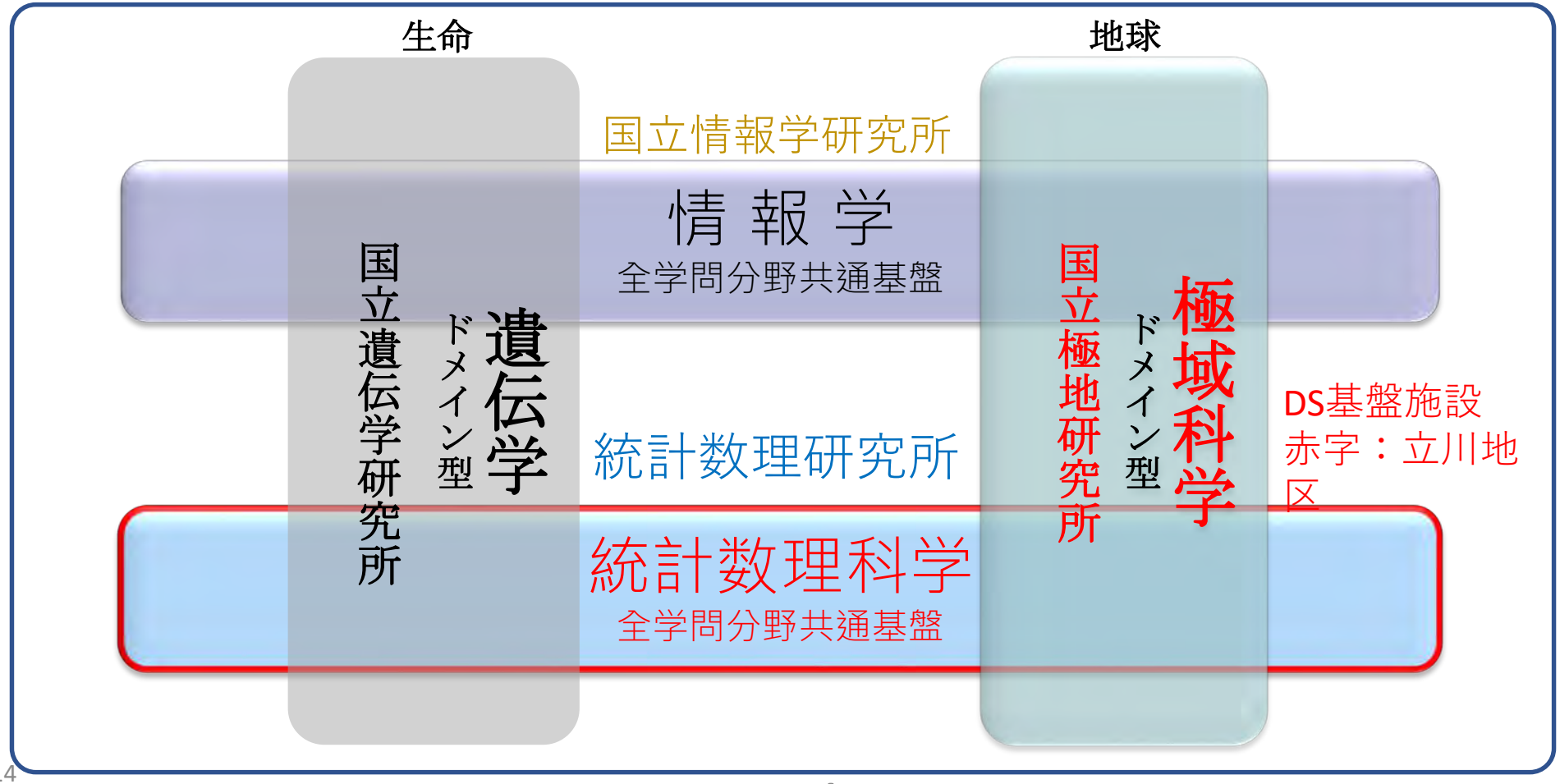
内容

- 自己紹介
- 私が受けた最初の統計の講義と行ったデータ分析
- 統計ソフトウェアと自身との関係性の変遷
- データ分析の曙
- データ分析に何を期待するのか？
- ソフトウェアで支援されたデータ分析教育の変遷
  - 特に：筑波大学大学院ビジネス科学研究科経営システム科学専攻
    - データ解析、多変量解析１，２、統計的管理、定量的研究の原理(調査票設計)
    - S-Plus⇒R+AMOS: データ分析vs実証研究
- 標準データセットとデータ分析コンペティション
- 統計的機械学習への途とこれからのデータ分析教育
- おわりに：普遍的プロセス

自己紹介

大学共同利用機関法人情報・システム研究機構

生命，地球，環境，社会などの複雑な問題（複雑科学）を，物質とエネルギーの観点に替って**情報とシステム**という立場から捉えるための，方法の研究，研究基盤の整備および融合研究による新分野の開拓を行なう．



【令和元（2019）年6月５日：創立昭和19年6月5日】

統計数理研究所　創立75周年

- 平成30（2018)年6月から「創立75周年記念事業」として
  オープンハウスをはじめとする各イベントを開催
- 75周年当日の令和元（2019）年6月5日に一橋講堂において
  記念式典を挙行

新発見とイノベーションへの飛翔

「研究力」「社会への還元・貢献」「異分野融合・新分野創成」*3*つのテーマの飛翔がもたらす、最先端のデータサイエンスにおける新発見とイノベーションを表現した記念ロゴを作成

自称古典統計家営業担当

- 東京大学工学部計数工学科助手：1982~1987
- **慶應義塾大学理工学部数理科学科講師：1987~1997**
- **筑波大学ビジネス科学研究科助教授・教授：1997~2011**
  - **ビジネス科学研究科・国際経営プロフェッショナル専攻・専門職大学院認定機関立ち上げ**
  - **ビジネス科学への統計や統計的機械学習の応用：1997-**
    - **マーケティング・ファイナンスなどの実証研究支援**
- 統計数理研究所リスク解析戦略研究センター創設：2005~2015
- 応用統計学会長、統計関連学会連合理事長
- 情報・システム研究機構理事・統計数理研究所長：2019~
- 横断型基幹科学技術研究団体連合副会長:2018~
- 生物統計・環境統計: 1981-
  - 医薬品許認可：中央薬事審議会新薬調査会
    - 臨床試験統計解析ガイドライン
    - 新薬第2調査会：循環器領域：市販後調査:メバロチン
  - 薬害肝炎訴訟原告側証人
- 環境計測: 1982-
  - 国立環境研究所衛星観測チーム
    - オゾン層・温暖化ガスの全球観測
  - 中央環境審議会]PM2.5環境基準・計測管理
- 公的統計:1990-
  - (独）統計センター理事長：2015~2019
  - 統計審議会・統計委員会
    - 調査技術開発部会・農林水産統計部会長
    - 匿名データ部会長：2019~:統計委員会委員長代理
- 品質マネジメント：1980-
  - 国際標準化：統計的方法の適用
  - ISO TC 69 ->207-> 176-> 69
    - Uncertainty of Measurement
    - Design for Environment: ISO Guide 64
    - 8 Quality Management Principles:SO 9000
    - Random Variate Generation:: ISO 28640:2010
    - Process based Quality Function Deployment
      - ISO 16355シリーズ
  - TQC企業指導: 1987-1997
  - (公財）日本適合性認定協会監理パネル:2003~2019
  - (一社）日本品質管理学会: 1980~
    - 統計・データの質マネジメント研究会
    - サービスのQ研究会立ち上げ
    - 製造業のためのビッグデータ解析研究会
    - 品質工学会と商品開発プロセス研究会
    - 日本品質管理学会長：2015~2017
    - 品質工学会特別顧問：2018~
- **統計教育と普及**
  - **産業界との協業**
    - **日科技連MA研、規格協会QRG**
    - **品質管理検定の立ち上げ**
  - **初中等数学・統計教育: 2005-**
    - **教育課程部会：算数・数学指導要領**
  - **内閣府数理・データサイエンス・AI教育プログラム認定制度検討委員会:: 2019~**
  - **ヘルスデータサイエンティスト協会理事・データサイエンティスト協会顧問**

さて統計家とは？：米国労働統計2017/05

Statistician

米国では統計家を政府がカウント：2018年から統計家＋数学家として

- 求人状況：2017年現在全米で就労者数36,540名，今後10年で33％増加（数学家と込み）と予測,

統計家は専門職として政府が統計上で定義！

- 統計的方法を用いてデータを収集・分析し，ビジネス，エンジニアリング，科学分野など
  実世界の問題のソリューション提供を支援する：Functional Managerと考えられる！

統計家となるための大学・大学院教育課程が成立

- 統計か，数学の修士学位が通常必要,
  最近のジョブ増加に伴い，学部レベルの職も増加．修士博士修了生は,
  手法開発やコンサルタントとして自立
- 統計学科修了生でも，統計家ではなく，アナリスト（金融アナリスト）や
  データサイエンティストと呼ばれる職種につく人もいる．

私が受けた最初の統計の講義と行ったデータ分析

田口玄一先生の統計学の講義のインパクト
データ分析（単回帰分析）を電卓で行っていた時代

1975年4月東大教養学部202教室「統計学」

- 田口玄一先生は，1975年4月から1年間東京大学教養学部理科Ⅰ類において，統計学の講義を担当されました
- 私は、その講義の受講生でした
  - 教科書は、丸善から出版された「統計解析」でした
  - 当時，田口先生は，青山学院大学に所属されていました．
  - 朗々とした口調で，
    - 「いいですか，良い技術者というものは，xxxのように考えるのです．」
    - 「ベル研では，yyy」
    - 「プリンストン大学のTukeyは、zzz」
  - といった声が朝の教室に響きました．
  - 講義内容もさることながら，高校まではエンジニアリング・センスなどという言葉に全く疎かったので，新鮮でした．
- 私は，その講義を受講し，はじめて統計解析という分野に興味を持ちました．
  - 特に分散分析，回帰分析，寄与率の計算の虜になりました．
  - この夏休みに，統計学を学べる学科はどこかということを調べ，計数工学科という学科に統計学の講座があることを知り，そこに進学することを決意しました．

1975年4月23日の講義「工学を横断する工学」

- R. A. Fisher 実験統計学
  - 技術情報の獲得効率を高めるための共通技術の全体
  - 実験計画とデータ解析
    - 行列，2次形式，直交展開（分散分析，直交多項式）
    - 統計ではランクの代わりに自由度という用語を用いる
- N. Wiener サイバネティクス
  - 通信と制御との共通理論
- C. Shanon ：通信理論
  - 情報の伝達効率を上げるための共通技術の全体
  - 情報量を数学的に定義

秋休みの宿題

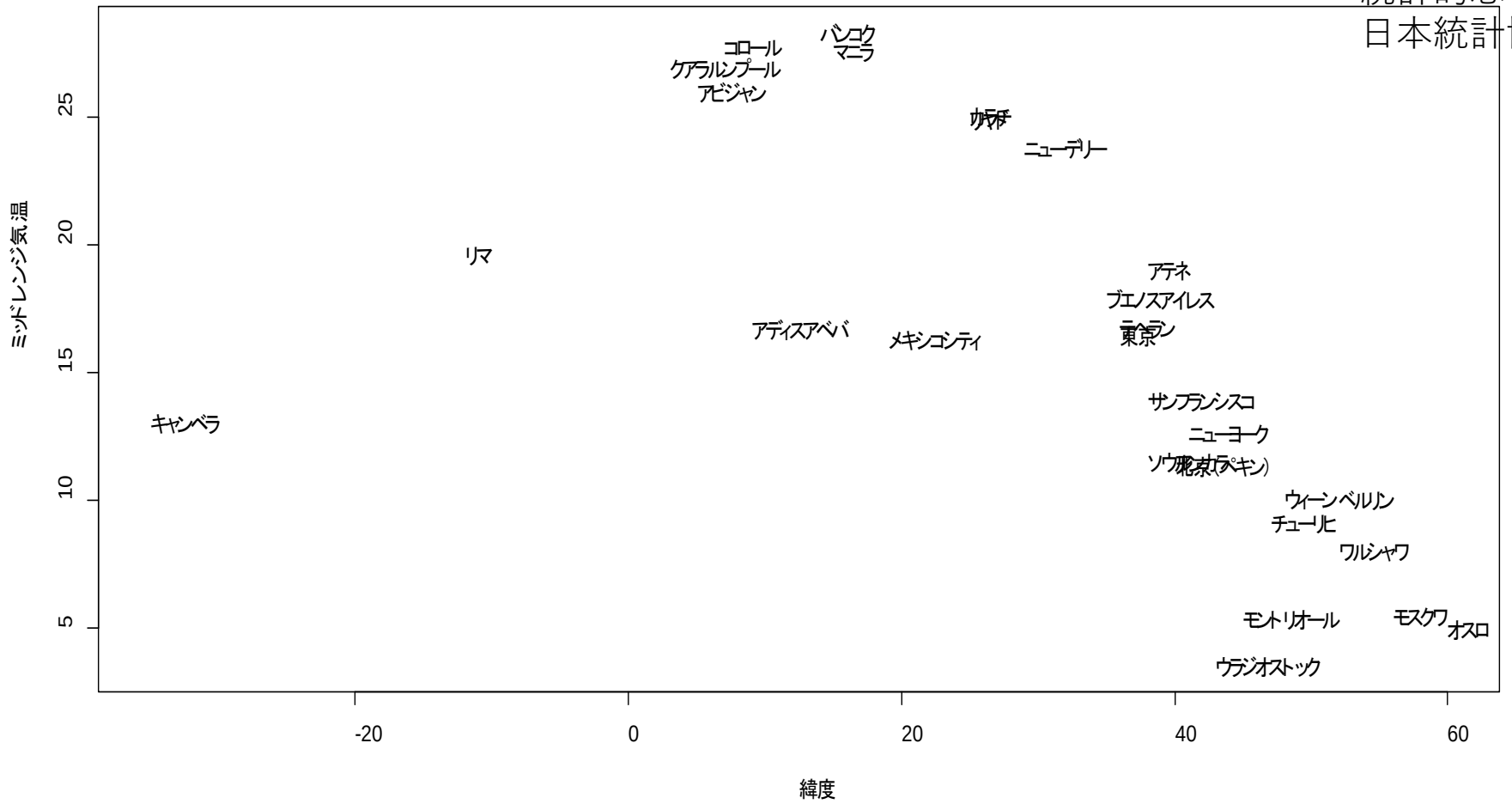
- レポート
- 自分又は他人による実験又は調査データの解析
  - **評価の視点**
    - **テーマの選定能力**
    - **解析能力**
    - **結論の出し方**
  - データの回帰，平方根変換，対数変換
- 自然地理・理科年鑑などの資料を基に，
  地球上の各地の月別平均気温と緯度（緯度の余弦変換）についての
  回帰分析を実施し，田口先生に提出しました.
  - 自分自身で合理的経験法則を作り出せることに感激.
  - 後年、田口先生と矢野宏先生との対話本に、この学生のレポートのことが記載されており嬉しかったです

後期1975年11月頃開始

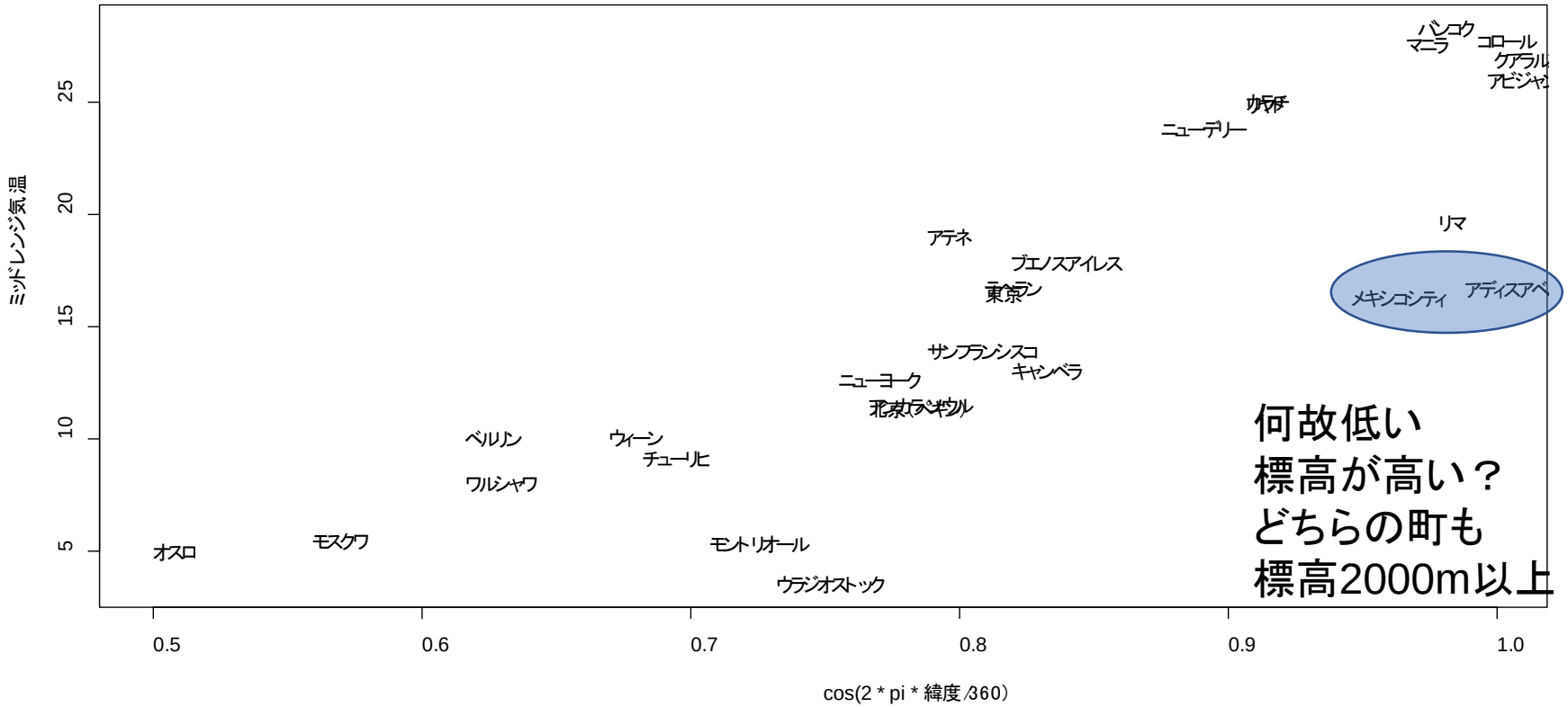
- 先生の講義では，１１月に入り，累積法などが教えられました
  - 何がなんだか分からず，後期期末試験では失敗したと思いました.
- **研究とは**
  - **目的：個性・着想**
  - **手段：技術力・創造力**
  - **評価：実験の計画・確実性・能率**
- **関数形による研究技法の分類**
  - **開発調査研究**
    - 関数形がある領域で不明：調査・実験計画
  - **理論の評価**
    - 関数形は予想がついているが未知母数がある：回帰分析
  - **設計計算**
    - 関数形が完全に所与：数値解析，摂動解析
      - SN比の考え方です。今日の数値実験計画法を予見されていました

北緯と気温との関係（理科年表）
科学的法則の認識
各月平均の最大値と最小値の平均
どんな関数形?→帰納から演繹への質問

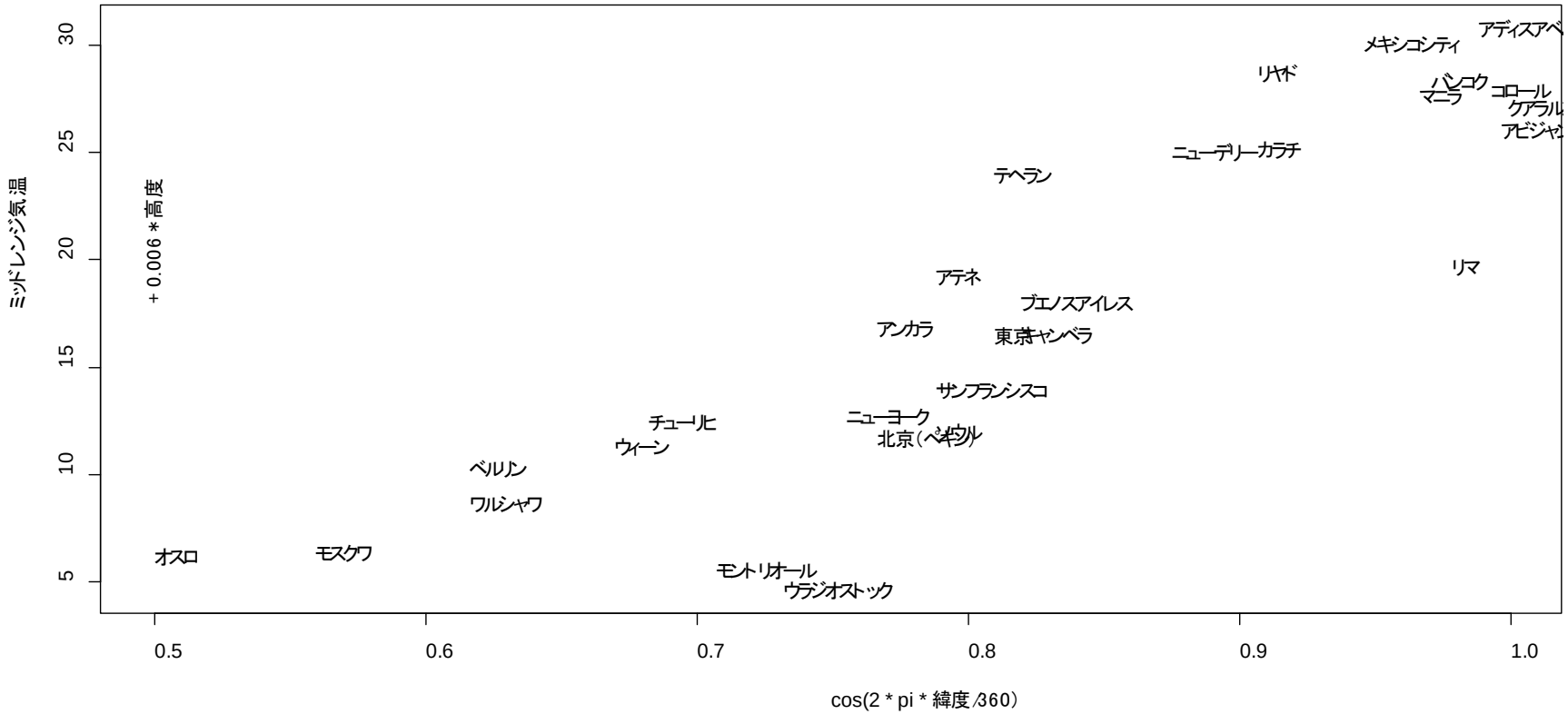
高校生用教材案
総務省政策統括官編
大学での学びにつながる
統計で身近な現象や社会の課題を探究するスタディガイド
高校からの統計・データサイエンス活用～発展編～
統計的思考力を身につけよう！
日本統計協会：2017年版



cos(緯度)と気温との関係：演繹
相関は0.86

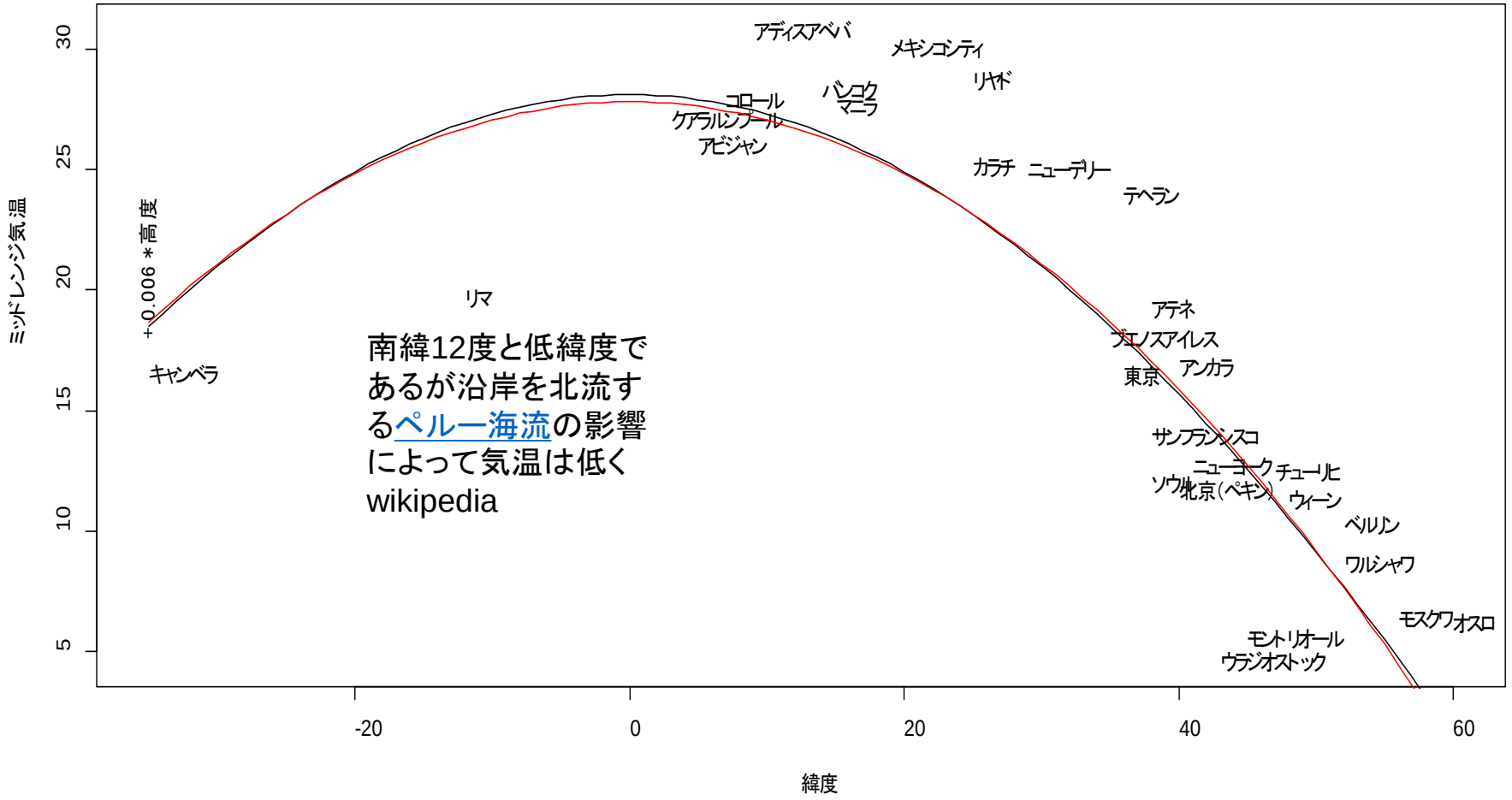


cos 緯度と気温+標高×0.006との関係：演繹
100m標高が上がると0.6度気温は下がる
相関=0.89



演繹と帰納とのすり合わせ：三角関数（黒）と2次関数（赤）

-25+53*cos(緯度) vs. 28-0.0075緯度^2



統計ソフトウェアと自身との関係性の変遷

統計プログラムパッケージ研究会⇒ＳＡＳユーザー会
Ｓの登場からＲへ
共分散構造分析やグラフィカルモデリングのツール
日本の統計ソフトウェアと産業界

私の統計分析ソフトウェアとの出会い

- 1979 教育用計算機センター・プログラマブル電卓(TI 59)
- 1981 電電公社回線網上の統計ソフトウェア群
- 1981 東大工学部計数工学科：統計プログラムパッケージ研究会
  - 大橋靖雄氏主催：SAS, BMDPの多変量解析方法を事例で紹介
  - 日本ＳＡＳユーザー会に発展
- 1982~2000?　日本科学技術連盟多変量解析研究会：探索的データ解析、GLIM, CARTなどをいち早く産業界に紹介
  - 奥野忠一、芳賀敏郎らのツール群（1975年頃）
  - 芳賀のCDA(Conversational Data Analysis)⇒JUSE-MA⇒ JUSE-STATWORK
- 1987~1997　慶應義塾大学理工学部数理科学科統計学コース
  - 渋谷政昭、柴田里程：S-Plus （ツール開発言語）による教育実践
- 1987 日本規格協会ポケットコンピュータによる品質管理研究会
  - 西堀栄三郎：Sharp PC 1501によるBasicプログラム開発
- 1995~2000 慶應義塾大学SFC：JMPによるデータサイエンス教育
  - 慶應SFCデータ分析教育グループ編「データ分析入門」への講義ノート提供
- 1996　日本品質管理学会テクノメトリックス研究会
  - 芳賀敏郎、廣野元久：グラフィカルモデル：GGM
- **1999~2011 筑波大学大学院経営システム科学専攻でのＲとAMOSによる教育**

印象的な活動

- 統計プログラムパッケージ研究会⇒ＳＡＳユーザー会
  - BMDP判別分析機能、SAS非線形最小二乗法機能
- Ｓの登場からＲへ
  - 大学におけるソフトウェアによる講義開始
- 共分散構造分析やグラフィカルモデリングのツール
  - 回帰分析から因果分析へ
- 日本の統計ソフトウェアと産業界
  - 日本科学技術連盟多変量解析研究会・日本規格協会データ解析研究会
  - 西堀栄三郎先生とポケットコンピュータ
  - 日本科学技術連盟品質管理教育へのデータ分析導入：JUSE STATWORK

自身の統計教育観を変えた文献

- 田口玄一(1972) 改訂新版統計解析,丸善
- Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method、Biometrika, Vol.61(3), pp.439-447.
- **奥野忠一**, **芳賀敏郎**, **矢島敬二**, **奥野千恵子**, **橋本茂司**, **古川陽子**(1976)続多変量解析, 日科技連.
- **芳賀敏郎**, **橋本茂司**(1980) 統計解析プログラム講座 1 統計解析プログラムの基礎,日科技連.
- Breiman, L., Friedman, J., Stone, C. J. (1984) Classification and Regression Tree, Chapman and Hall
- **高橋行雄**, **大橋靖雄**, **芳賀敏郎**(1989)SASによる実験データの解析, 東大出版会
- Bollen, K.(1989) Chapter 6: Measurement Models: The Relation between Latent and Observed Variables, in *Structural Equations with Latent Variables* ,John Wiley & Sons.
- 三浦新編(1989)ポケコングラフィックス活用による品質管理的アプローチ, 日本規格協会. (西堀栄三郎先生の目指したコト)
- Chambers, J. M. and Hastie, T.J. (1991) Statistical Models in S, Chapman and Hall, 柴田里程訳(1994), Sと統計モデルーデータ解析の新しい波, 共立出版
- 狩野裕(1997) AMOS、EQS、LISRELによるグラフィカル多変量解析—目で見る共分散構造分析,現代数学社.
- 日本品質管理学会テクノメトリックス研究会編(1999)グラフィカルモデリングの実際, 朝倉書店.

データ分析の曙

データ解析の登場

日科技連多変量解析研究会

Tukeyとデータ分析（解析）
Future of Data Analysis & EDA

- データ分析誕生
  - Tukey, 1962
- 裁判官の立場の推測統計
  - 仮説検証型統計推測
- 刑事の立場のデータ分析
  - Tukey, 1977
- Bell 研究所の指導
  - C言語からS言語へ
    - データ分析環境

J.W. Tukey, J.H. Friedman and M.A. Fisherkeller
Stanford Linear Accelerator (1973)
26:00 minutes

http://stat-graphics.org/movies/prim9.html

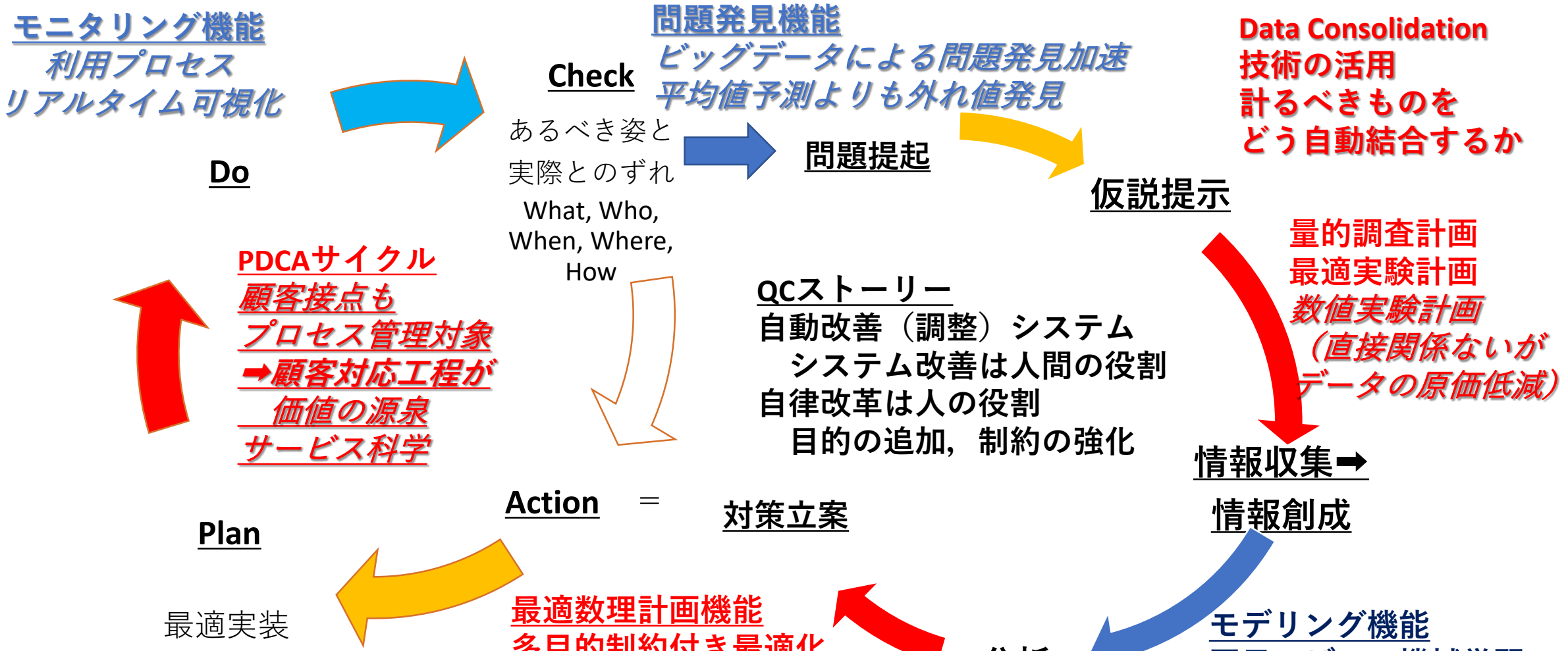
データ分析に
何を期待するのか？

目的の分類：発見・分析・最適化

モデリング（分析機能）のプロセス

何のための数理・データサイエンス・AI教育

デミング・石川モデルのSHINKAの方向性
データサイエンスによる問題解決



データ分析は科学の文法に組み込まれる

- Shewhartの問題発見
  - 外れ値の発見と要因分析
  - 本質的な外れ値とデータ分析の未熟さによる外れ値
- 分析のプロセス
  - 研究すべき問題，仮説の定式化
  - 関連するデータの探索と適切なデータを採取する研究の計画と実施
  - データ解析
  - 適切な意思決定に繋がる結果の解釈
- 最適化
  - 前工程の分析結果が最適化に寄与しやすい出力か？
    - 応答曲面法（2次形式の最大化）
  - Cox & Donnely(2011) Principles of Applied Statistics, Cambridge University PressのIdeal Processも認識科学的だがプロセスモデル字

経営系専門職大学院QAプロジェクト 2005年度ー2006年度

- ビジネススクールの教育の質保証システム検討会
  - 国内経営系大学院（ＭＢＡ相当の学位を授与する大学院）をメンバーとする検討グループを設立
  - ビジネススクールの「教育の質」と「修了生の専門職としての質」を保証する新たなシステムの制度設計
- 2006年1月13日：ビジネススクール長会議で合意

プロジェクトの内容

① Plan（教育目標設定）: **必要な経営専門職とは、どのようなコンピテンシーを有する人材か**：筑波大学

② Do（教育実施方法）: それを系統的に育成するビジネススクールの教育プログラムには、どのような要件が求められているのか：青山学院大学

③ Check（教育評価方法）: ビジネススクールにおける教育の質を保証するシステムとは、どのようなものであるか：同志社大学

筑波大学大学院ビジネス科学研究科担当力量指針のScope

- 規定内容
  - ビジネススクールが育成すべき人材，すなわち経営専門職候補者の力量について規定
- 対象
  - ビジネススクールを設置している、あるいはこれから設置しようとしている大学の関係者
- 狙い（Purpose)
  - ビジネススクールの教育システム設計に資する共通基盤を与える
  - ビジネススクールが自律的に設計する教育システムの妥当性を検証可能なものにする
- 備考　（Note)
  - 各ビジネススクール独自の教育目標並びに教育システムを束縛することは意図していない
  - ビジネススクールで習得すべき具体的経営分野の知識・技能並びにその水準を規定することを意図していない。

筑波大学国際経営プロフェッショナル専攻が起案したビジネススクールで育成すべき経営専門職の力量

- 経営専門職のマネジャー行動を支える力量
  - 3段階のマネジャー行動：**統計で支援できる行動もある！**
    - ***価値*と問題の発見**
    - **意思決定**
    - 適用・実現
  - グローバルに有用な10の力量（Competency)
    - 永井裕久，椿広計(2005)「筑波大学の新ビジネススクール：デザインから誕生までを振り返る」，経営行動学, 18(2), 145-153
  - 永井裕久教授らによる調査研究
    - 製造業(自動車、電機・電子、化学、石油)におけるグローバル企業20社のミドルマネジャー約2000名とその上司を対象に質問票調査
      - 渡邊寿美子，永井裕久，河合忠彦，田代美智子(2004)高業績グローバルマネジャーのコンピテンシー活用に関する国際比較調査,国際ビジネス研究学会年報 (10), 201-215
- **統計的接近によって高まる力量も多々ある！**

			経営専門知識（提供科目群体系）									
		専門知識１次	組織経営			事業戦略				応用情報		
経営プロフェッショナル力量		専門知識2	経済理論	ガバナンスと倫理	組織行動	質マネジメント	コストマネジメント	金融	マーケティング	統計分析	OR	情報システム
力量要求一次	力量要求二次	力量重要度										
問題発見	多様性受容	10		1	3	1			3			1
	達成志向	10	1		3	3	1	1				
	先見性	10	1					3	3	1	1	1
意思決定	情報収集力	8					1	1	1	2		2
	創造性指向	8			2	1		1				1
	分析思考	8	2				2	1	2	2	1	
	戦略立案	8	1			1		1	2	1	2	
	リスクマネジメント	8		2	1		1	1		1	2	
適用・実現	組織マネジメント	15		2	4	4	2	2				
	コミュニケーション	15		4	2	2			2			4
		知識寄与度	5	9	15	12	7	11	13	7	6	9

行動	コンピテンシー	行動特性
問題発見	1) 多様性受容　Ability to accept diversity	異なる視点を検討し、異なる意見を傾聴することにより、様々な可能性を考慮に入れる技能
	2)達成志向 Commitment to sucesss	不確実性が高く、解決が困難な状況において、課題達成の手段を探索する技能
	3) 先見性　Ability to anticipate problem	**解決すべき課題に影響を与える現在および、将来の諸要因を見通す技能**
意思決定	**4) 情報収集力　Ability to gather information**	**意思決定に必要な質的に高い情報を効率的に収集する技能**
	5)創造性志向 Creative Thinking	既存の概念を組合わせたり、新たな発想にもとづいて課題に取り組む技能
	6) 分析思考 Analytical orientation	**課題解決に適合的な情報と手法を選択して分析する技能**
	7) 戦略立案 Strategic plannning	複数の評価尺度の検討から、高い成果が期待される施策を作成する技能
	8) リスクマネジメント　Risk management	**客観的にリスク発生の確度およびその影響を把握し、発生した場合の対処における役割が担当できる技能**
適用・実現	9) 組織マネジメント Organizational management	与えられた経営資源の配分やメンバーの意識に配慮し、統括部門の目標を達成する技能
	10）コミュニケーション Communication skills	意思疎通における曖昧な状況を排除するとともに、関係者から支持や理解を得る技能

渡邊，永井他(2004)に見る力量と実務実績との関係

	利益獲得	コスト低減	付加価値	信頼関係構築	品質改善	プロセス改善	新ビジネス創生	後継者育成	組織ノウハウ蓄積	上級管理職への可能性	昇進可能性	平均達成水準の向上
経験からの学習力	欧州 日本	中国	世界共通								欧州	世界共通
不確かさマネジメ			世界共通	中国			アジア	中国		中国		アジア
結果とプロセスの管理											中国	
変革推進		中国		米国	日本			中国		中国		
コンフリクトマネジメント			日本									
誠実さ	欧州		アジア								アジア	世界共通
ネットワーキング			世界共通									
プレゼンテーション					世界共通		日本					

	利益獲得	コスト低減	付加価値	信頼関係構築	品質改善	プロセス改善	新ビジネス創生	後継者育成	組織ノウハウ蓄積	上級管理職への可能性	昇進可能性	平均達成水準の向上
多様性受容			世界共通		世界共通							アジア
達成志向	欧州	米国	世界共通							アジア	アジア 中国	
先見性	日本		日本		日本		世界共通	世界共通				
情報収集力			アジア				世界共通	欧州	欧州	世界共通	欧州	
創造性志向	欧州	世界共通										
分析思考			世界共通							欧州 日本	欧州	
戦略立案	日本											
リスクマネジメント					日本							
組織マネジメント				世界共通			中国					
コミュニケーション	世界共通		世界共通	世界共通						日本		日本

ABEST21第2次調査(2009)
力量開発に適した教育方法
N=386
日本・環太平洋ビジネススクール関係者

Competencies	Training Methods						
	OJT	Virtual Project	Group Discuss.	Case Studies	Lecture	Reading	NA
Ability to accept diversity	44.0	42.0	**64.0**	36.7	18.7	12.7	5.3
Ability to learn from experience	**68.0**	38.0	27.3	48.0	17.3	12.0	2.7
Commitment to success	**46.0**	42.7	20.0	19.3	16.7	9.3	11.3
Managing uncertainty	40.0	39.3	18.7	**46.0**	24.0	12.7	6.0
Ability to anticipate problems	25.3	30.0	19.3	**40.7**	25.3	21.3	16.7
Ability to gather information	**36.7**	32.0	31.3	30.7	34.7	26.7	8.0
Creative thinking	24.0	**44.0**	32.0	32.0	20.0	12.0	18.7
Analytical orientation	23.3	46.7	31.3	**57.3**	46.0	32.7	2.7
Strategic planning	32.0	**58.7**	32.7	54.7	39.3	31.3	3.3
Risk management	34.0	34.0	24.7	**46.0**	33.3	22.7	8.7
Organization management	42.0	36.0	36.7	**44.0**	32.0	23.3	4.0
Result Process management	**46.0**	39.3	24.7	38.0	25.3	13.3	9.3
Change agency	**48.0**	38.7	22.7	32.7	13.3	6.7	15.3
International Communication	**37.3**	23.3	31.3	12.7	26.7	20.0	20.7
Domestic Communication	42.0	28.0	**46.0**	14.0	17.3	14.0	12.7
Conflict management	44.0	48.7	**50.7**	24.0	10.0	4.0	5.3
Integrity	**40.0**	17.3	24.0	14.0	8.7	9.3	35.3
Networking	**44.7**	36.0	38.7	10.0	7.3	5.3	18.0
Presentation skills	36.7	**53.3**	50.0	30.0	30.7	6.7	4.7

力量開発に適した
知識

Competencies	Technical Expertise						
	Finance Account	Market Strategy	Corp. Gov.	HRM	Q&O Manage	IT & Data A	Ethics SR
Ability to accept diversity	10.1	53.0	21.5	**59.7**	10.1	6.7	41.6
Ability to learn from experience	32.2	**59.7**	18.8	37.6	32.2	17.5	28.9
Commitment to success	26.2	**57.1**	12.1	24.2	20.8	18.8	10.1
Managing uncertainty	**45.0**	44.3	25.5	18.1	22.2	31.5	17.5
Ability to anticipate problems	29.5	**73.2**	11.4	12.1	16.1	36.9	9.4
Ability to gather information	25.5	47.0	9.4	13.4	17.5	**75.8**	10.1
Creative thinking	8.8	**77.0**	8.8	12.2	17.6	17.6	9.5
Analytical orientation	56.8	49.3	8.8	12.8	26.4	**66.2**	5.4
Strategic planning	41.9	**83.8**	21.6	23.0	20.3	33.8	12.8
Risk management	**50.7**	39.2	36.5	18.2	23.7	27.7	33.8
Organization management	10.1	39.9	29.1	**72.3**	21.0	7.4	21.0
Result Process management	24.3	**50.0**	18.9	21.0	41.9	27.7	11.5
Change agency	17.6	**58.8**	18.2	24.3	21.0	23.7	12.2
International Communication	12.2	**29.1**	12.8	25.7	10.8	8.8	10.8
Domestic Communication	13.5	**36.5**	12.8	30.4	12.8	12.2	14.2
Conflict management	10.8	31.1	18.2	**54.7**	16.2	8.8	14.9
Integrity	8.8	13.5	29.1	33.8	11.5	6.1	**59.5**
Networking	6.1	37.2	15.5	**41.2**	15.5	18.9	12.8
Presentation skills	20.3	**55.4**	14.9	17.6	22.3	23.0	12.2

ソフトウェアで支援されたデータ分析教育の変遷

統計ソフトウェアを教室で使うことを前提としなかった時代の教育
統計ソフトウェアを誰もが家で使える前の統計教育
統計ソフトウェアを誰もが使えるようになった後の教育

統計ソフトウェアを教室で使うことを前提としなかった時代の教育

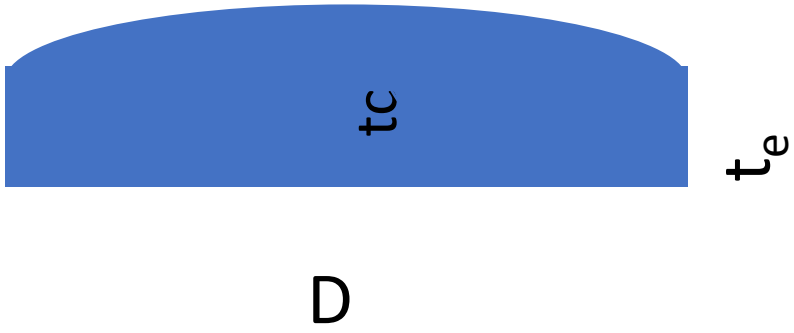
ソフトウェアの背後にある事例教育

川崎浩二郎先生（コニカ）との共同
日科技連Basic Courseへの重回帰教育導入

- 仮想ストーリー
  - プラスチックレンズの成形歪みは、ニュートン本数を計測することによって評価される。
  - 一般に成形後は型に対して凹の方向、ニュートン本数にしてプラスの方向に歪みが大きくなるので、型は一般にマイナス側に設計してある。
    - 型の概要は次に示す。
  - 加工後の歪み量には、多くの要因（代用特性）が影響していると考えられる。この要因を探るために、様々なレンズについてその形状条件（Ｄ，ｔｃ，ｔｅ）、及び加工条件（成形材質、成形圧力、成形温度）のデータを採取した。
  - 与えられた、Ｄ，ｔｃに対して、歪み量の平均値がニュートン本数にして－７となるような条件を探索せよ。ちなみに、規格限界は－７±６である。
    - 但し、０．５≦ｔｅ≦１．５，成形圧力は２．５（kg/cm^2）以下，成形温度は１２０度以下に押さえないと成形内部歪みが大きくなってしまう（圧力、温度は低めに押さえられるならそれにこしたことはない）。
    - また、素材１は素材０に比べてかなりコスト高になってしまう

プラスチックレンズ成形歪み計測データ

代用特性						
形状条件			加工条件			品質特性
D	ｔｃ	ｔｅ	材質	圧力	温度	ニュートン本数
15	2	0.5	0	2.0	118	-7
15	2	0.5	0	2.1	111	2
15	2	0.5	1	2.2	103	2
15	2	1.0	0	2.1	106	-2
15	2	1.5	1	2.2	102	-8
15	3	0.5	0	2.2	100	15
15	3	1.0	1	2.2	101	-2
15	3	1.5	1	2.1	105	-7
20	3	0.5	1	2.0	120	-4
20	3	1.0	0	2.2	112	-1
20	3	1.0	0	2.3	101	9
20	3	1.5	1	2.3	100	-4
20	4	0.5	1	2.2	109	14
20	4	1.0	1	2.3	101	7
20	4	1.5	1	2.4	100	-4
20	4	1.5	0	2.4	102	3
25	3	0.5	1	2.2	115	9
25	3	1.0	0	2.3	106	10
25	3	1.5	0	2.3	101	9
25	3	1.5	1	2.3	112	-8
25	4	1.5	1	2.1	118	-9
25	4	1.5	0	2.5	103	5
25	5	1.0	1	2.2	118	0
25	5	1.5	1	2.4	117	-7
25	5	1.5	1	2.5	105	-2



但し、
材質０は標準素材、材質１は新素材

相関係数行列

	N	D	TC	TE	TR	DR	PLAS	PRESS	TEMP
N	1.00000	0.09044	0.03322	-0.44832*	0.52977**	0.03023	-0.40318*	0.19568	-0.35353
D	0.09044	1.00000	0.67337**	0.39863*	0.00493	-0.04402	0.13878	0.53446**	0.29835
TC	0.03322	0.67337	1.00000	0.41231*	0.16941	0.69348**	0.33821	0.59817**	0.13483
TE	-0.44832	0.39863	0.41231	1.00000	-0.77283**	0.18196	0.15677	0.59339**	-0.28921
TR	0.52977	0.00493	0.16941	-0.77283	1.00000	0.23334	0.14169	-0.27564	0.36447
DR	0.03023	-0.04402	0.69348	0.18196	0.23334	1.00000	0.31961	0.29817	-0.17042
PLAS	-0.40318	0.13878	0.33821	0.15677	0.14169	0.31961	1.00000	0.00000	0.17380
PRESS	0.19568	0.53446	0.59817	0.59339	-0.27564	0.29817	0.00000	1.00000	-0.47867*
TEMP	-0.35353	0.29835	0.13483	-0.28921	0.36447	-0.17042	0.17380	-0.47867	1.00000

諸代用特性の中で品質特性に対して最も相関の強いのは、
teの-0.45ということになる。しかし、既にお気づきのように上の行列には、
初期のデータ表には、存在しないｔｒ＝ｔｃ／ｔｅという代用特性が存在している。これは、固有技術者の成形歪みは、「レンズの厚みが一様でないから生じる」といった意見を勘案して追加した合成特性である。
実は、これが一番相関が強い有用な代用特性となったのである。
現実問題では、このような固有技術の助けによって相関の高い代用特性が発見されることが多いし、望ましいことでもある。

課題４：１節の仮想ストーリで単相関分析の結果は満足のいくものか検討し、更なる改善の可能性を探れ。

現状把握：単回帰・相関分析によれば、特性ｔｒのコントロールが最も有用となった。そのときの単回帰式は、

$$N = -7.56 + 2.38 \times tr \qquad r=0.530$$

となっている。従って、Ｎのねらい値を－７にするためには、ｔｒは0.235程度をねらい値とする必要があるが、ｔｃ≧ｔｅであるから、ｔｒは、１以上にしかできない。
仮にｔｒ＝２としても,Ｎのねらい値はＮ＝－２．８と全く不満足なものである。また、分散もｔｒ値のバラツキを完全に０にしても、

$7.234^2 \times (1-0.530^2) = 39.28 = 6.268^2$,　6.268は改善後の品質特性の標準偏差

依然として満足できない。更なる改善活動は必須である。

改善活動１：仮に、ｔｒ＝2.0に標準化した上で、再びデータを採取し、単回帰、単相関分析を繰り返す。→可能ならば、大変結構なことです（本手）。

改善活動２：　現有データでなんとかならないか？
重回帰的シミュレーションの前提：ｔｒは品質特性のみならず全ての代用特性に影響を与えている。その影響は単回帰分析で評価できる。実際のシミュレーションは例えば、以下のように行う。

仮に、ｔｒ＝2.0 に調整した後に、ｔｅとＮの相関がどうなるかをシミュレーションしてみよう。
ｔｅは、ｔｒの調整前には、Ｎに対する単相関がｔｒに次いで大きかったからである。

ｉ番目のデータの品質特性、$N_i = -7.56+2.38\,Tr_i + e_i$：ｅは近似誤差（残差）

ｉ番目のデータのｔｅ: $te_i = 1{,}797-0.204\,Tr_i + f_i$：$f_i$ は近似誤差（残差）

すると、ｔｒ＝2.0に制御した後のデータは次のように修正される。

ｉ番目のデータの第１次改善後品質特性推定：

$$N_i = -7.56+2.38x2.0 + e_i = -2.8 + e_i$$

ｉ番目のデータの第１次改善後のｔｅの推定値: $te_i = 1{,}797-0.204x2.0 + f_i$

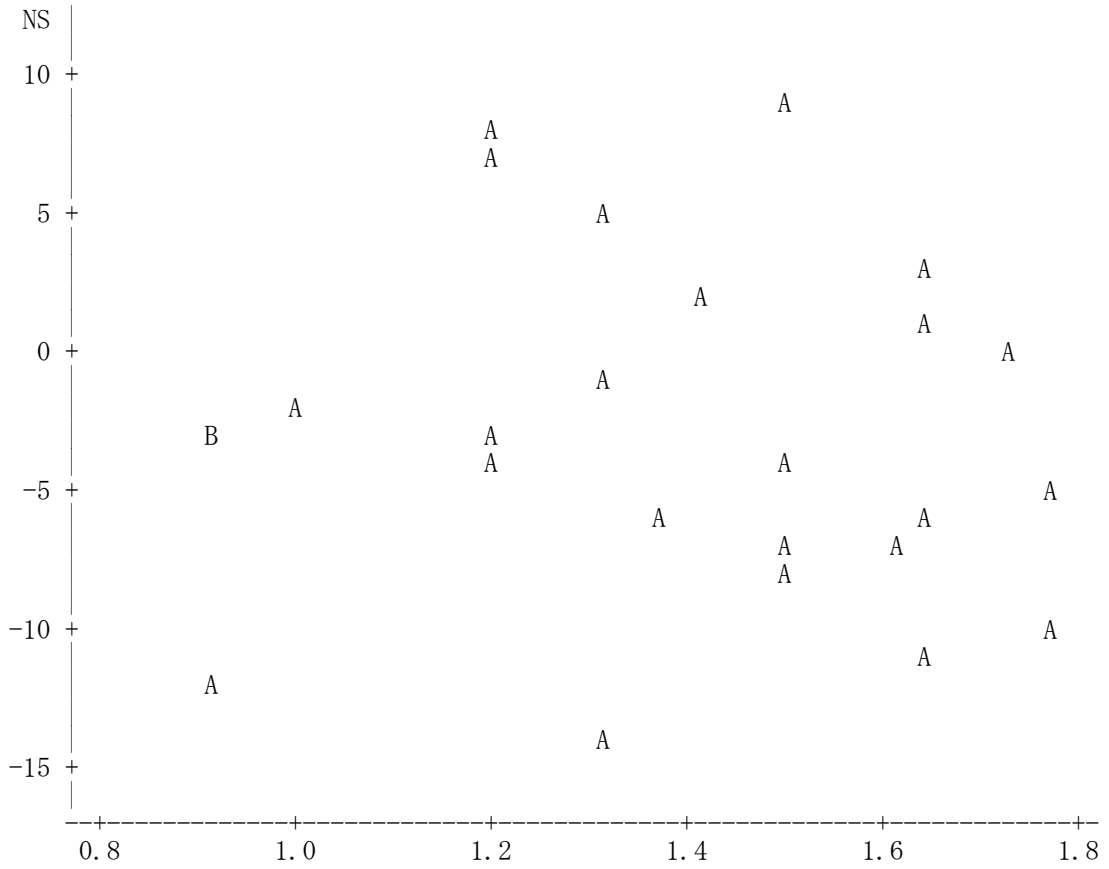
$$= 1.389 + f_i$$

改善後のシミュレーションデータ

OBS	N(Simluate)	Te(Simulate)	N	Te	N	Te
1	-11.7761	0.90985	11	6.6079	1.20542	21	-10.5975	1.63728
2	-2.7761	0.90985	12	-4.0082	1.50099	22	3.4025	1.63728
3	-2.7761	0.90985	13	-0.3118	1.72756	23	-7.1600	1.61428
4	-2.0082	1.00099	14	2.2239	1.40985	24	-10.1868	1.77357
5	-6.4189	1.36471	15	-5.5975	1.63728	25	-5.1868	1.77357
6	5.4561	1.31871	16	1.4025	1.63728			
7	-4.3921	1.20542	17	-0.5439	1.31871			
8	-7.0082	1.50099	18	7.6079	1.20542			
9	13.5439	1.31871	19	8.9918	1.50099			
10	-3.3921	1.20542	20	-8.0082	1.50099			

改善シミュレーションの散布図

プロット ：NS*TES.　凡例 ： A =



N＝-0.517-1.643 t e

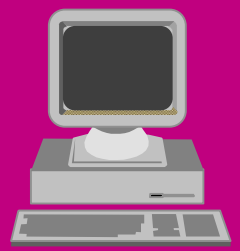
統計ソフトウェアを誰もが家で使える前の統計教育

Sによる教育開始

単純なデータ（選択制）による授業内演習

データ解析同演習１９９４－１０－１７
椿　広計（注:慶應義塾大学理工学部数理科学科）

データセットの作成と Ｓへの読み込み



計算機室での集合教育
単純なデータで定型的課題提示

用意された原データ（１）

è１）世界の国の諸指標
è２）日本の都道府県の諸指標
è３）業界別経理指標
è４）各社の就職条件
è５）世界の気候

用意された原データ（２）

è６）　東横線住宅価格
è７）　食品成分表
è８）　自動車カタログ
è９）　飛行機カタログ
è１０）　プロ野球（１９９３）成績
èその他各自が用意できるデータ

１）世界の国の諸指標
「世界の統計１９９４」より

- 国内、国民総生産
- 名目、実質成長率
- 国民所得
- 支出項目別、経済活動別国内総生産及び実質成長率

- 総資本形成、部門別貯蓄
- 輸出入総額
- 貿易指数
- その他

２）都道府県諸指標
「日本の統計１９９４」より

- 県民経済計算指標
- 普通会計歳入歳出額
- 地方交付税交付額
- 行政投資実額
- 消費者物価地域指数
- 住宅地平均価格

- 医療施設数、医療関係者数
- 学校数、教員数、在学者数、進学率、就職率、教育費
- 交通事故

３） 業界別経理指標

- 損益計算
  - 売上高、売り上げ原価、売り上げ総利益
  - 販売費、営業利益、経常利益、特別利益
  - 特別損失、法人税、中間配当金
- 貸借対象表
  - 流動資産、固定資産、総資産、資本
- 参考：割引手形など
- 安全性諸比率

４）各社就職条件

- 資本金、売上高、社員数、年齢
- 役員数など
- 大学初任給
- 勤務条件
- 採用実績
- 休暇、福利厚生、諸制度

５）世界の気候：理科年表

è月別平年降水量
è月別平均気温
è月別相対湿度

６）東横線沿線マンション価格
週間住宅情報９４／１０／１２

- 駅名
- バス、徒歩
- 価格、坪単価
- 専有面積
- バルコニー面積
- 間取り情報

- 総戸数、向き、階
- 築年月、駐車場
- 管理費、管理形態
- 特記事項（公庫融資有無、角部屋など）

７）野菜の栄養：食品成分表

- 廃棄率、エネルギー
- 水分、蛋白質、脂質
- 脂肪酸、コレステロール
- 炭水化物、食物繊維、灰分
- 無機質、食塩相当量
- ビタミン

８）自動車カタログ

- 社名
- ボディ形式
  - ドア数、シート数
- エンジン
  - 総排気量、シリンダ数
  - 最高出力、最大トルク
  - その他

- シャシー
- ブレーキ、ステアリング、タイヤ
- 車両寸法
- 性能
- 価格

９）飛行機カタログ

- 全幅、全長、全高、主翼面積
- 最大離陸重量、機体重量
- 最大巡航速度、最大航続距離
- 離陸滑走距離
- 運行乗員数

１０）プロ野球成績’９３

- 打撃成績
  - 試合、打数、得点、安打、二塁打、三塁打、本塁打、塁打、打点、盗塁、盗塁刺、犠打、犠飛、四死球、三振、併殺打、打率、失策他
- 投手成績
  - 試合、完投、無点勝、無四球、勝利、敗北、セーブ、勝率、打者、投球回、安打、本塁打、四球、死球、三振、失点、自責点、防御率

行列型データファイルの作成

è１）データ行列とは
  –nxp行列に数値を配列
  –各行は、個体に対応
  –各列は、変量に対応
è２）ﾃﾞｰﾀﾌｧｲﾙとは
  –数値変数、文字変数を
  ｎ行ｐ列に配置したテキ
  ストファイル

```
12.7  258  0.23  yes
13.2  301  0.78  no
.....................
11.2  298  0.97  yes
```

Sへのデータの取り込み(1)

作成された test.dat

```
2.57 12.3 352 yes a
3.14 13.2 217 no b
2.83 20.0 256 yes c
3.25 17.2 289 no a
3.42 18.2 215 no a
2.71 17.9 3 yes c
```

Sへのデータ取り込み（２）

Ｓでの取り込み

```
 template<-list(0,0,0,"","")
> scan(file="c:test.dat",template)
```

```
> scale(test.x)
      length     weight      width
a  1.5786571 -1.4555365 -1.37565255
b -1.1737986 -1.1326534  0.49767635
c -0.3786447  1.3069078 -0.52115165
d  0.2941778  0.3023826  0.85919596
e -1.2145757  0.6611416  1.41790809
f  0.5592291  0.5535139 -0.91553668
g  0.3349549 -0.2357559  0.03756048
>  test.comp<-prcomp(scale(test.x))
```

```
> test.comp
$sdev:
[1] 1.3880742 0.9128352 0.4898796
$rotation:
           [,1]       [,2]       [,3]
[1,] -0.6591333 -0.2140799 0.72091131
[2,]  0.3961853 -0.9136582 0.09091691
[3,]  0.6392030  0.3455408 0.68703784
$x:
         [,1]       [,2]        [,3]
a -2.4965289  0.5165599  0.06061356
b  0.6430653  1.4581123 -0.60725954
c  0.4342333 -1.2930859 -0.51220014
d  0.4750978 -0.0423646  0.82986794
e  1.9688330  0.1459040  0.15866410
f -0.7345263 -0.9417975 -0.17552999
g -0.2901742  0.1566719  0.24584407
```

課題２－１

- 本日紹介したデータを各自分業で適当に意味のあるデータとして入力し、データファイルを作成する。
- データファイルをＳの中に取り込み、ＳのＬＩＳＴオブジェクトを生成し、変数名を付ける。数値変量については、平均、標準偏差を算出してみよう。
- データ行列も作成してみよう。

統計ソフトウェアを誰もが使えるようになった後の教育

プロジェクトベースの学習によるGood Practice の共有

統計計算法教育から統計の基本原理の教育へ

慶應義塾大学SFCのデータ分析教育開始

- SFCでデータ解析の講義担当：当初S言語
  - 400名のデータ分析のレポート
  - 紙ヘリコプター実験
- 1990年代後半：データサイエンスを標榜！
- 入学生全員にJMP inプレインストールされたPC
  - 当時SAS社の岸本淳司先生（現在九州大学）の尽力
- 「データ分析」
  - 1年生にJMPによる記述統計、重回帰分析、主成分分析
  - データソース、データの品質概念教育：椿の講義ノート提供
  - 実際のデータ分析（データ取得）
  - データ分析結果プレゼンテーションによるGood Practice共有
- 理屈は興味のあった人間に教える「統計解析」
  - 尤度概念・確率モデル・統計的推論

筑波大学大学院ビジネス科学研究科　多変量解析第一

各自のPC上のRによる予測モデル仮説成長型データ解析教育

上場企業2091社100変量最大10年分

3名から4名のプロジェクトのプレゼンテーション＋個人の第2課題レポート

多変量解析（第１回）
講義の狙いと概要
データ解析の流れと仮説の成長

１９９８－０９－０２

筑波大学大学院経営システム科学
椿 広計（つばき ひろえ）

GSSMでの統計教育

- 夏休み：データ解析
  - 基礎概念、記述統計、線形モデル、記述多変量
- 2学期：多変量解析
  - モデルの構築と成長
    - 樹形モデル、加法モデル、一般線形モデル
    - グラフィカルモデル、パス解析、因子分析入門
- 3学期：統計モデル
    - 検証的方法：尤度と仮説検定
    - 統計モデルにおける潜在因子の役割
  - 調査設計、博士課程講義、輪講、計量経済学、時系列モデルで他のニーズに応える
- 注）2004年頃から「統計的管理」を追加
  - 検査（意思決定）、異常発見、実験計画をテーマ

講義計画：戦術(1)

- ９月２日：オリエンテーション：
  - 講義の狙いと概要
    - １時限 多変量解析の戦術と戦略
    - ２時限 班分け：
      データ配布とSplusの基本的使い方
- ９月９日：曖昧な仮説から定性的知識へ
    - １時限 ：樹形モデルからの出発
      - データに潜む論理構造
    - ２時限：班別演習

講義計画（２）：戦術(2)

- 9月16日：知識の定量パターン化
  - 1時限：論理構造から非線型ﾊﾟﾀｰﾝ構造への進化
    - 一般加法モデルの当てはめ
  - 2時限：班別演習
- 9月30日：定量パターンの計量モデル化
  - 1時限：パターンの数式化
    - 一般線形モデルの当てはめ
  - 2時限：班別演習
- 10月７日：1,2時限 第1回班別発表会と討論

講義計画（３）：戦略(1)

- 10月14日：検証的因果分析入門
  - 1時限：単一方程式から連立方程式へ
    - 連関図の定量化
  - 2時限：班別演習（AMOSデモ版配布）
- 10月21日：探索的因果分析
  - 1時限：グラフィカル・モデリング入門
    - 偏相関の意味
  - 2時限：班別演習（CGGMの使い方）
- 10月28日：探索的因子分析
  - 1時限：潜在因子の役割と単純構造
  - 2時限：班別演習（Splus)

講義計画（４）

- 11月4日：因子分析＋因果分析＝共分散構造分析
  - 1時限：再び連関図のモデル化：様々な事例
  - 2時限：計算機演習（AMOS)
- 11月11日：1,2時限：第2回班別発表会と総合質疑

  - この他、多変量解析輪講第1、第2を必要に応じて、2学期から春休みにかけて開講します。第1は、「講義」の方法の理論的側面に興味のある方、第2は、授業で取り上げない統計手法を追加学習したい方のために、土曜日などを活用して適宜ゼミ形式で開講いたします。
    受講を希望される方やグループは、10月までにご連絡下さい。

1.データ解析とは何か？
1.1 データの採取と解析

- データ解析(Data analysis)とは？
  - 情報に付加価値を生じさせる一連のマネジメント・プロセスを扱っており、料理と類似
- データ解析の目的
  - 科学的説明(Scientific explanation)
    - 科学的仮説の探索:exploratory study
    - 科学的仮説の検証:confirmatory study
  - 技術評価(Technology assessment)

1.2 統計的データ解析の発展と今日的意義

- この四半世紀の主要な改善
  - 一般化線形モデル
    - ロジスティックモデル、比例ハザードモデル等
      - 目的変数の多様性への配慮
  - 線形潜在構造モデル
    - 確証的因子分析、共分散構造分析
      - 非観測変数が本質的役割を持つことへの配慮
  - モデル診断手法の生成
    - 線形性の欠如、外れ値、交互作用、その他

多変量データ解析の潮流

- 検証的解析：多変量仮説検定
  - T.W.Anderson(1958)
    - An Introduction to Multivariate Statistical Analysis, Wiley.
- スローガンとしての探索的データ解析
  - J.W. Tukey(1977)
    - Exploratory Data Analysis, Addison Wesley.
- モデル接近への回帰？ 椿の思い込み？
- 新しいパラダイム：折衷的解析
  - 純粋な探索的データ解析から「曖昧な仮説」の仮説成長型データ解析へ

データ解析で
期待されるモデル達(戦術)

- 樹形モデル
  - CART, AID
  - データ間構造を
    if-then ruleで表現
  - 情報量的に最適な
    Decision Treeを推定
  - 数式的関係の表現は
    苦手だが、データの
    構造を最も柔軟に
    探索

- 一般加法モデル
  - 線形モデル

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

  - 加法モデル

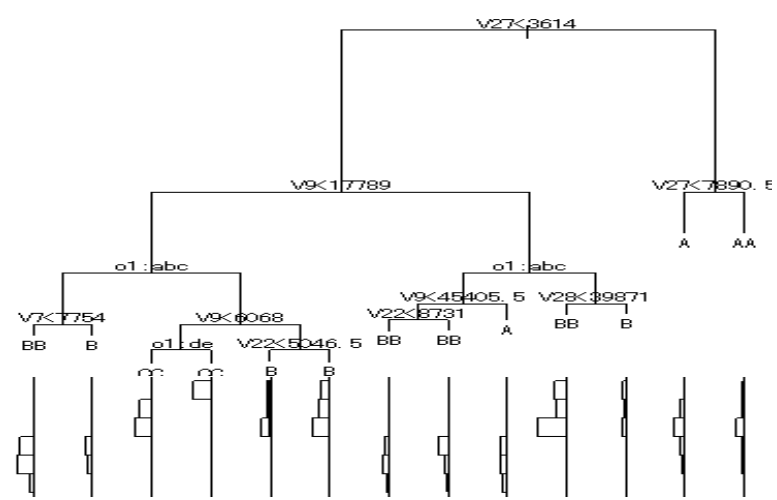
$$\lambda = \backslash(x) + \varepsilon$$

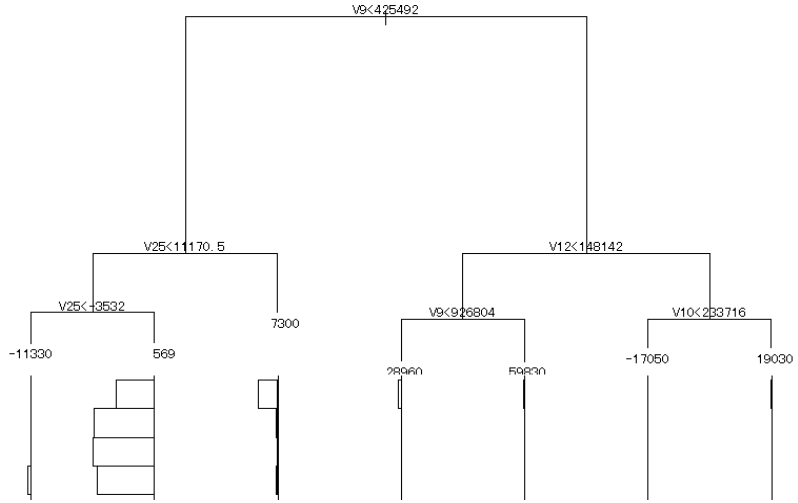
    関数形を指定しない
    曲線パターンが出力

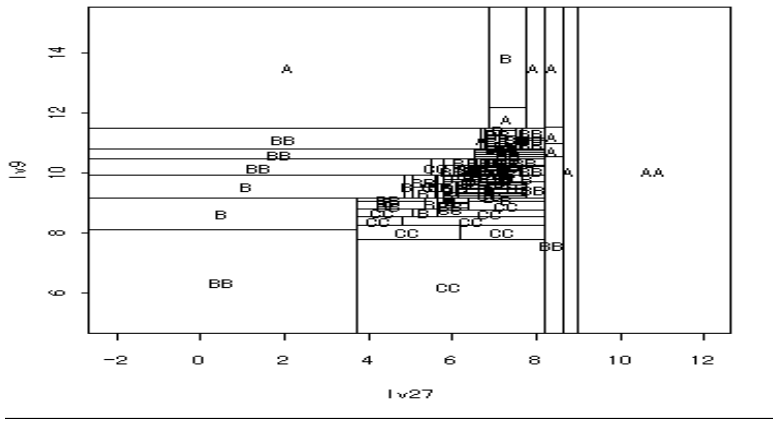
1.3仮説成長型データ分析の手順と基礎概念

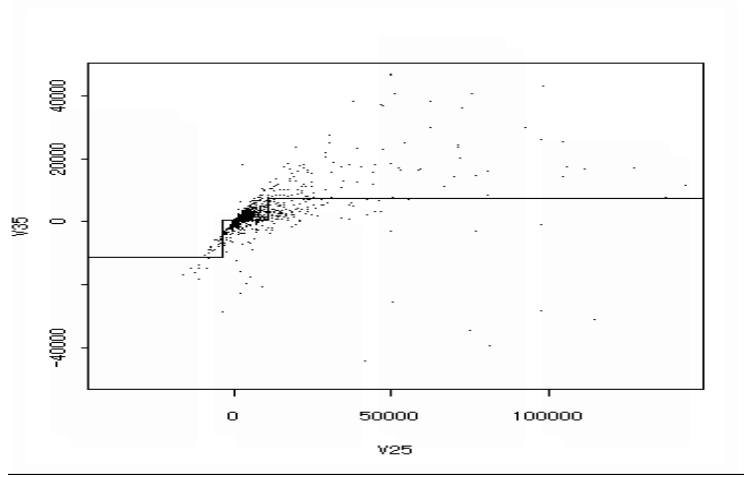
- 起)曖昧な仮説の提示（原始仮説）
- 承)樹形モデル当てはめで
  if-then規則型仮説への成長
  - 交互作用、例外構造の探索
- 転)一般加法モデル当てはめで
  パターン型定量仮説への成長
  - 非線形構造の探索
- 結)数式モデル(一般線形モデル等）の
  当てはめで、メカニズム型仮説への成長

承：If-then ruleからの出発



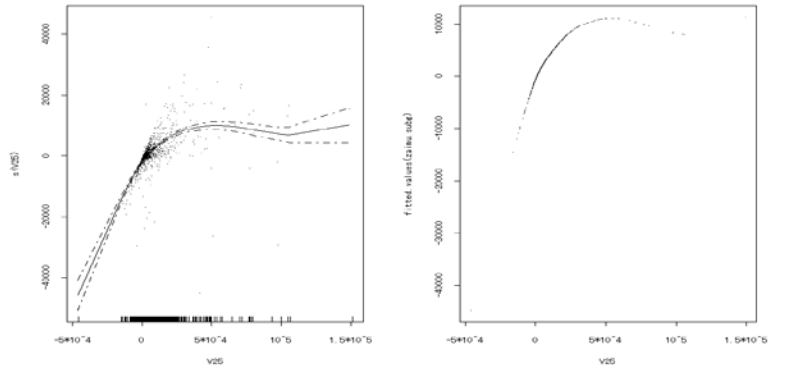




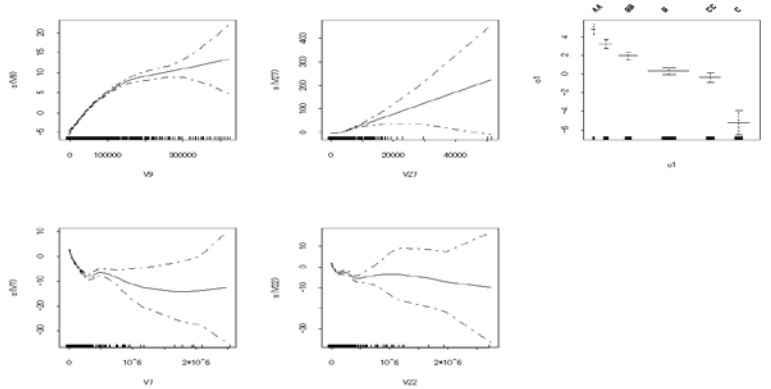


転：要因効果のパターン抽出

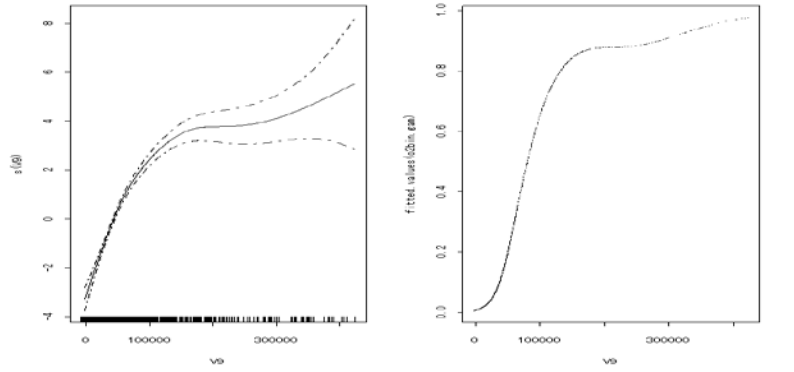
散布図平滑化



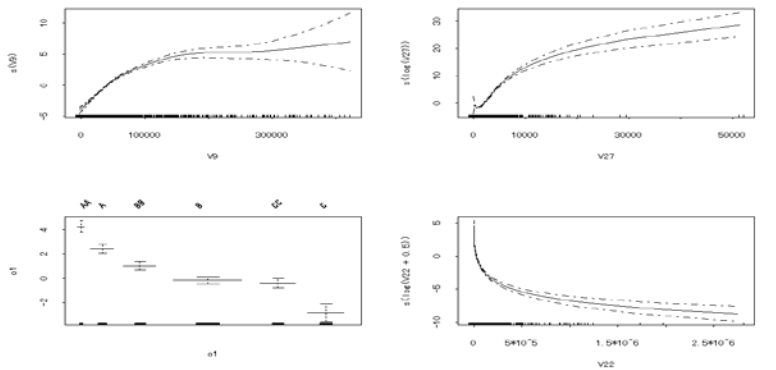
安全性への要因効果



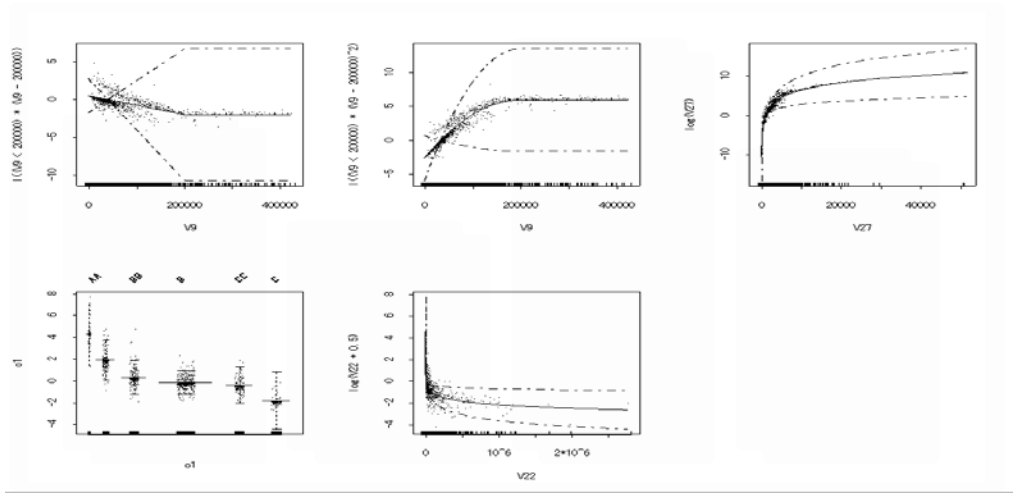
質的予測の場合



変数変換による安定化



結：要因効果の数式化



- >glm(formula = o2.bin ~ I((V9 < 200000) * (V9 - 200000)) + I((V9 < 200000) *(V9 - 200000)^2) + log(V27) + o1 + log(V22 + 0.5), family = binomial)
- ここには、固有科学の知識が必要

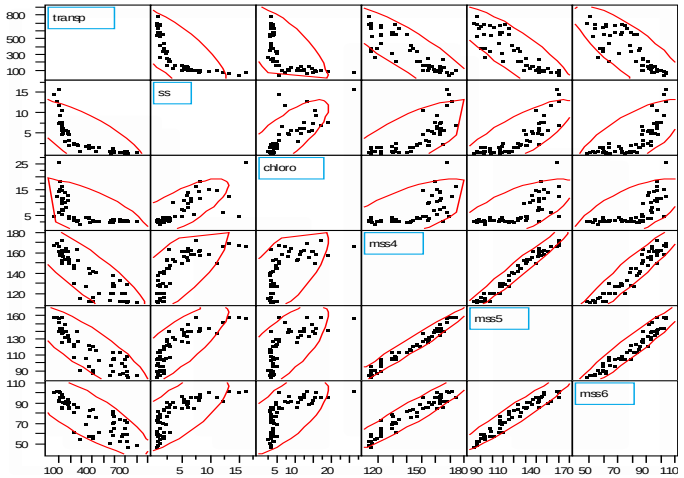
データ解析戦術としての統計工学の危機？

- 情報採取技術（実験・調査計画）の隆盛
- データ解析技術としての統計利用の衰退？
  - Camelon(1997, ISI review)の指摘
    - 既に統計が、この分野を独占していると思うな!
  - 技術目的（予測、決定）の最適化のツールとしては古典的統計手法よりは、情報工学分野の手法が既に優れている。
    - Computer intensive statistical methodの多くは、同等だが、情報工学分野との発想の境界は希薄となっている
      - 注：1998年頃の椿の見解ですが

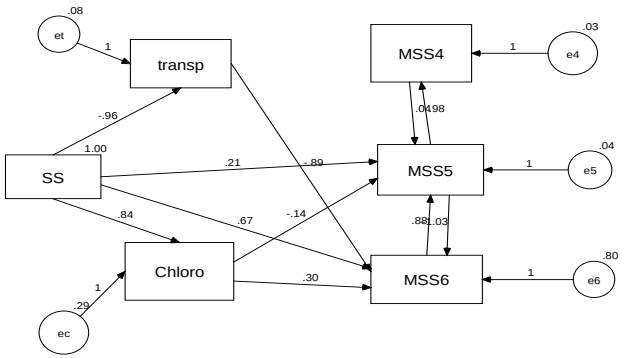
1.4 Explanatory Analysis
モデルビルディング(戦略論)

- 現象のメカニズム推論の意義
  - 統計科学へ：Fisherの科学的推論への回帰
    - 統計により、与えられる付加価値とは、現象に対するより定量的な理解
  - データの変動を「説明」するモデルを当てはめ、作業仮説を成長させよ！
    - モデルの要素として以下が有ることを認識せよ
    - 構造： 普遍的構造部分
    - 測定： 技術的限界が突破されれば変動しうる構造
    - 標示： 採取したデータに特有の構造
  - 標示･測定部分に含まれる母数は、モデル利用の場に応じて再チューニングせよ
  - 一般化（外挿）可能性への挑戦こそ、モデル思考の意義

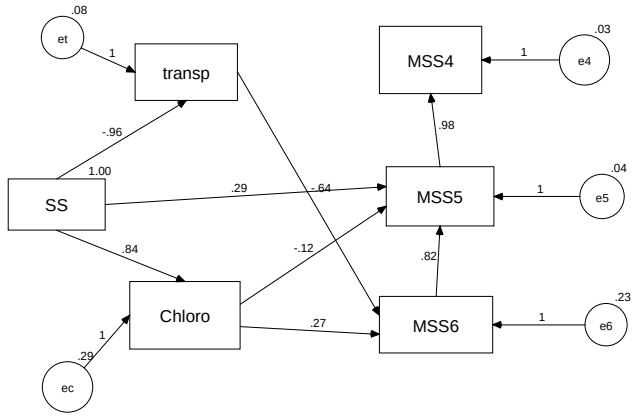
琵琶湖環境計測モデルの進化



1. Raw data



2. After G.M.



3. After Cov. Structure Analysis

$$\Sigma = \begin{bmatrix} \sigma_{SS}^2 & -0.96\sigma_{SS}^2 & 0.84\sigma_{SS}^2 & 0.19\sigma_{SS}^2 & 0.996\sigma_{SS}^2 \\ -0.96\sigma_{SS}^2 & 0.92\sigma_{SS}^2 + 0.08 & -0.81\sigma_{SS}^2 & -0.18\sigma_{SS}^2 & -0.96\sigma_{SS}^2 - 0.05 \\ 0.84\sigma_{SS}^2 & -0.81\sigma_{SS}^2 & 0.71\sigma_{SS}^2 + 0.29 & 0.16\sigma_{SS}^2 - 0.03 & 0.84\sigma_{SS}^2 + 0.05 \\ 0.19\sigma_{SS}^2 & -0.18\sigma_{SS}^2 & 0.16\sigma_{SS}^2 - 0.03 & 0.036\sigma_{SS}^2 + 0.044 & 0.19\sigma_{SS}^2 + 0.03 \\ 0.996\sigma_{SS}^2 & -0.96\sigma_{SS}^2 - 0.05 & 0.84\sigma_{SS}^2 + 0.05 & 0.19\sigma_{SS}^2 + 0.03 & 0.993\sigma_{SS}^2 + 0.29 \end{bmatrix}$$

4. Estimated Covariance

配布データ
東洋経済の財務データ96

- 86-96年の上場企業の財務諸表データ
  - cvsファイル３つ
- ファイル
  - 企業名と96年度の東洋経済による評価
  - 変数名（日本語）
  - ｘｘ年度財務データ
- 各班：2-3名で共有する事
- 今日の演習：各自データをSplusに読み込む
- データ・フレイムを作成し、次の命令を行う
  - >attach(ﾃﾞｰﾀﾌﾚｰﾑ)
  - >summary(データフレーム）
  - >plot(変数１，変数2)

２．データに潜む論理構造を明らかにする樹形モデル当てはめ

多変量解析第1第2講

2000年9月09日

Hiroe TSUBAKI

講義の概要

- １）層別とその最適化
- ２）自動層別
- ３）事例による樹形モデルの記述
- ４）計算機演習
  - 樹形モデル当てはめ
  - 枝刈り
  - 結果の表示あれこれ

１．層別(stratification)とその最適化

- 層別とは(旧JIS Z8101品質管理用語）
  - 母集団をいくつかの層(stratum)に分けること
    - 層別は，層内ができるだけ均一になるように，
      層間の差が大きくなるように行うと有利である.
  - 層別の良さを測るためには，層内の不均一性の尺度(impurity measure)が必要
    - 代表的なのがDeviance（尤離度，逸脱度，乖離度）
      - 数値変量→残差平方和
      - 因子変量→ｴﾝﾄﾛﾋﾟｰと同等

層別の最適化を少し詳しく

- Impurity measure
  - 目的変数が集団内で不均一であることを測る尺度
  - 目的変数が計量値：残差平方和

$$Se = \sum \left(y_i - \bar{y} \right)^2$$

  - 目的変数が質的変数：独立性のカイ二乗
    - エントロピー

$$D = -2 \sum_k n_k \log \hat{p}_k, \quad \hat{p}_k = n_k \Big/ \sum_{k'} n_{k'}$$

層別による集団の純化

- 性質
  - 層別を行う前の集団のimpurity measureは，
    層別後の各層のimpurity measureの総和より，必ず大きい
  - 層別を行うことは，よりPureにすること
  - 層別の効率とは？
    - impurityの減少が，最大となるような層別が良い層別！
      - 群間一様性検定統計量が指標となる
        - F値、カイ二乗値など

自動層別(AID）
Automatic Interaction Detector

- 説明変量の値（数値変量，順序尺度なら大小関係，分類尺度なら一致，包含関係）に基づいて，層別の効率が最大となるような反応変量の層別方法を自動探索
- Morgan and Sonquist(1963)Problems in the analysis of survey data, and a proposal. J.American Stat. Assoc.,58,415-434.

筑波大学ビジネス科学研究科
「統計的管理」

統計的決定理論：倒産確率推定と投資決定

統計的プロセス管理（異常値発見）

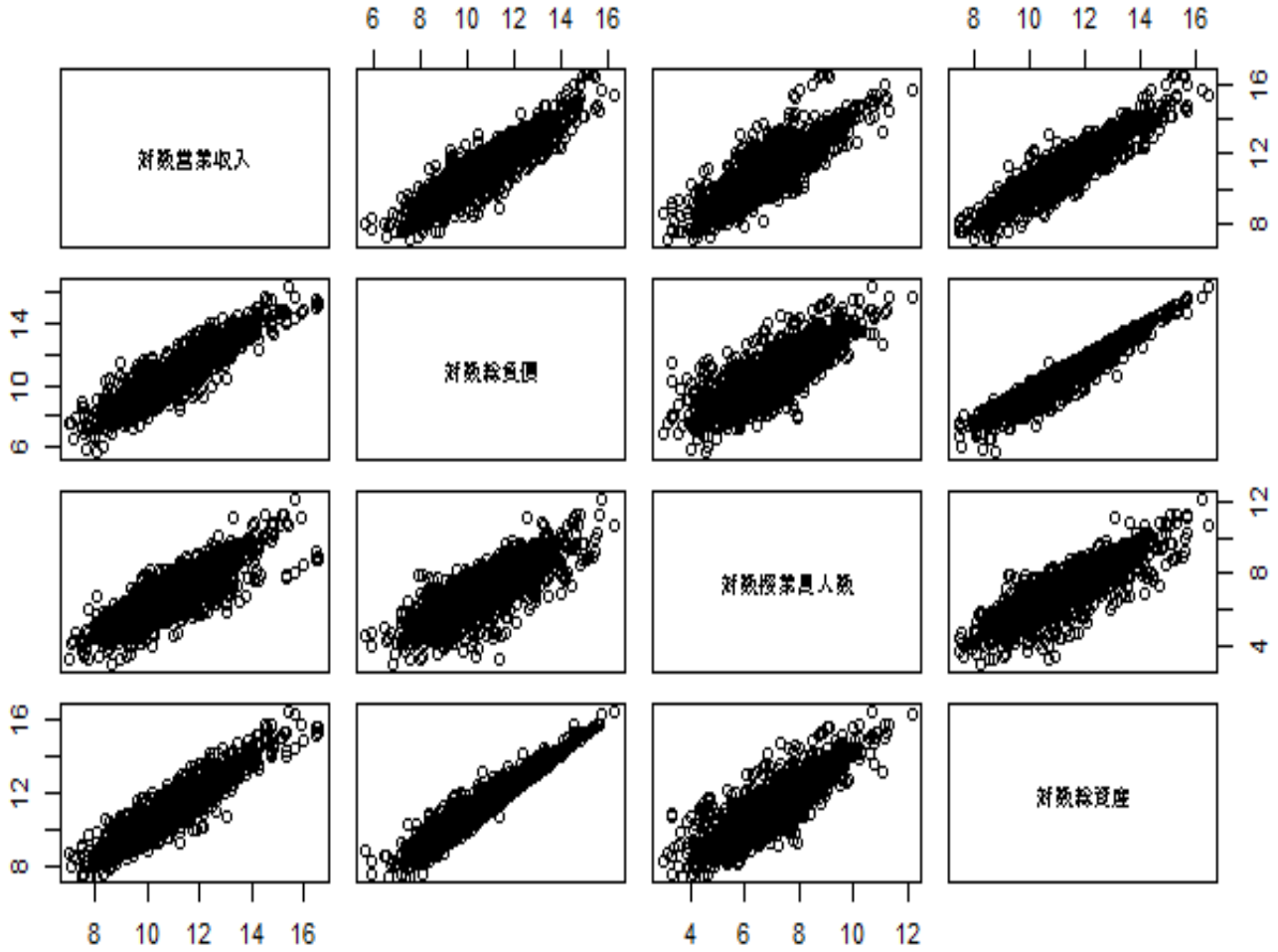
実験計画法：Conjoint分析によるビジネスモデル最適化レポート

多変量管理図法

- 多変量特性データへのプロセス管理の拡張
- 本質的に異なる２つのデータアプローチ
  - 「**識別分析 (Identification Analysis)**」：多変量管理図
    - Shewhart管理図を多変量特性に拡張
      - 多様な管理外れを検出
      - MT法
  - **統計モデルアプローチ**
    - 多変量データに入出力モデルを適合
    - 入出力関係の「外れ値」を検出

事例による比較
1996年度のわが国の上場企業2091社の4変量財務データ
東洋経済新報社財務カルテから抽出

- 対数営業収入
- 対数総負債
- 対数総資産
- 対数従業員人数

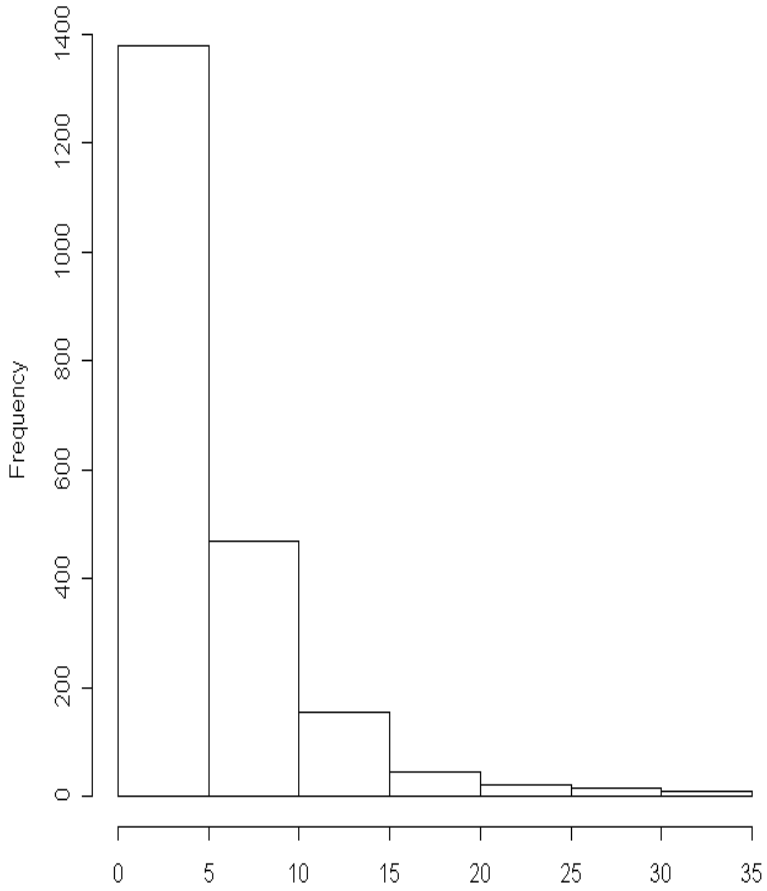


Flury and Ridwyl(1988)古典的識別分析

- 平均ベクトル$\boldsymbol{\mu}$
- 分散共分散行列$\boldsymbol{\Sigma}$既知のp変量正規特性$\mathbf{Y}$
- マハラノビス距離D^2
- $D^2=(\mathbf{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$ (1)
  - 多変量正規モデルの下でD^2は，自由度pのカイ二乗分布
  - $\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$の推定値を計算するための，統計的管理状態と見なせる標本が必要
  - MT法:正常群データから構成される線形空間を単位空間
- 単位空間に属するデータを事前に選別できるとは限らない
  - データ分析に用いる標本を統計的管理状態にあると仮想
  - 管理外れと見なされるデータを外れ値とみなす
  - 外れた原因を考察するというアプローチを行う必要性

識別分析：平均値、共分散
マハラノビス距離のヒストグラム

	対数営業収入	対数総負債	対数従業員人数	対数総資産
Mean	10.9	10.4	6.9	11.0
Cov.	1.95	1.90	1.45	1.78
	1.90	2.23	1.45	1.96
	1.45	1.45	1.50	1.41
	1.78	1.96	1.41	1.89



マハラノビス距離25以上
データ22社を外れ値として摘出

Chugai Mining	ORIX INTERIOR
MIYAKOSHI	FANUC
ITOCHU	Marubeni
SUMITOMO	TOKYO SANGYO
Sumitomo Realty & Development	HOKKAIDO SHINKO
East Japan Railway	SEA-COM
Fuji Kisen	Nihonbashi Warehouse
Tokyo Electric Power	Chubu Electric Power
Kansai Electric Power	Chugoku Electric Power
Tohoku Electric Power	Shikoku Electric Power
Kyushu Electric Power	Hokkaido Electric Power

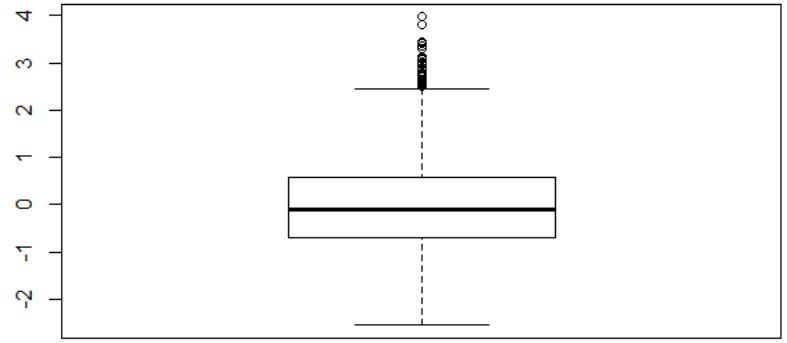
- 大手商社や電力会社が殆ど外れ値
  - データは統計的管理状態にはない？
  - 合理的な群分けを行う必要
  - これらの業種のデータを除いた標本を用いて識別分析を継続
- 外れ値の影響を受けにくい平均値ベクトルや共分散行列をロバスト推定
  - 外れ値を効果的に識別
- これらのアプローチでは大手商社や電力会社の単位空間を構成できない.
- この困難を回避
  - データ自体を多変量有限混合分布として表現
  - 確率的クラスタリング(潜在クラスモデリング)

入出力関係の外れ値抽出

- 入出力関係の外れ値を摘出
- その原因を追究
- 未然防止や予兆発見
  - 対数営業収入:出力変数
  - 出力に影響を与えている入力変数
    - 対数総負債
    - 対数従業員人数
    - 対数総資産
- 共分散行列を連立方程式モデルで表現
  - 完全逐次モデル（飽和モデル）豹変
  - 適切な構造モデルが想定される場
  - 構造モデルの残差に基づいて管理外れ摘出
  - より有用なプロセス管理情報

- 対数従業員数 $=\alpha_1+\beta_{12}$対数総資産$+\varepsilon_1$
- 対数総負債$=\alpha_2+\beta_{13}$対数総資産 $+\beta_{23}$対数総資産$+\varepsilon_2$
- 対数営業収入$=\alpha_3+\beta_{14}$対数総資産$+\beta_{24}$対数総資産 $+\beta_{34}$対数総資産$+\varepsilon_3$
  - E[対数]資産]$=\mu$,
  - Var[対数]資産]$= \sigma_0^2$
  - E[ε_j] =0,
  - Cov[ε_j, $\varepsilon_{j'}$]$=\delta_{jj'}\sigma_j^2$.

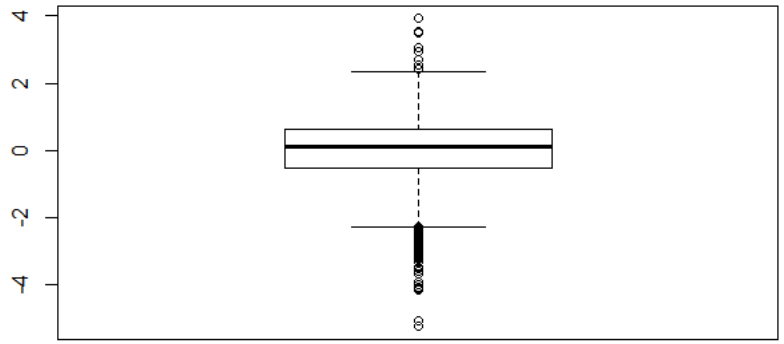
STEP 1　最上流「対数総資産」の外れ値摘出（識別分析）



μ+3σ以上のデータを管理外れと見なすと次の15企業が抽出
大手商社，電力会社はこの段階で
合理的群分けの対象
単に大企業が摘出されている可能性

- Hitachi, Matsushita Electric Industrial
- TOYOTA MOTOR, ITOCHU, Marubeni
- MITSUI & CO., SUMITOMO, Mitsubishi
- East Japan Railway,
- NIPPON TELEGRAPH AND TELEPHONE
- Tokyo Electric Power, Chubu Electric Power
- Kansai Electric Power, Tohoku Electric Power
- Kyushu Electric Power

STEP 2　入出力関係の外れ値摘出

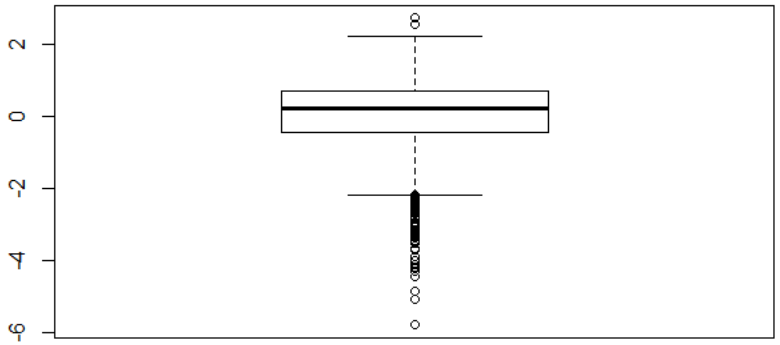


対数従業員数 $=\alpha_1+\beta_{12}$対数総資産$+\varepsilon_1$
をデータに当てはめ
その残差を標準偏差1に標準化したデータの
箱ひげ図
標準化残差2.5以上の企業8社
運輸業種を合理的群分けにすべきとする仮説が提示
Nippon Express,　YAMATO TRANSPORT
FUKUYAMA TRANSPORTING
Daiwa Motor Transportation
TOKYO BISO KOGYO NIPPON KANZAI
CENTRAL SECURITY PATROLSN TECHNO

- 標準化残差-3以下の企業31社
  不動産業が摘出

Chugai Mining,　Shoei,　ORIX INTERIOR
CAROLINA,　ATSUGI NYLON INDUSTRIAL
OKABE,　KOTOBUKI INDUSTRY, MIYAKOSHI
Toshoku, Footwork International,　Mitsui Fudosan
HEIWA REAL ESTATE, Tokyo Tatemono
DAIBIRU, SANKEI BUILDING, TOKYU LAND
HANKYU REALTY,　Keihanshin Real Estate
L Kakuei, Sumitomo Realty & Development,
Towa Real Estate Development,
TAIHEIYO KOUHATSU, NICHIMO, TOC
URBAN LIFE, ANA REAL ESTAT
ISUZU CONSTRUCTION, AIRPORT FACILITIES
Mitsui O.S.K.Lines,　HINODE KISEN, SEA-COM

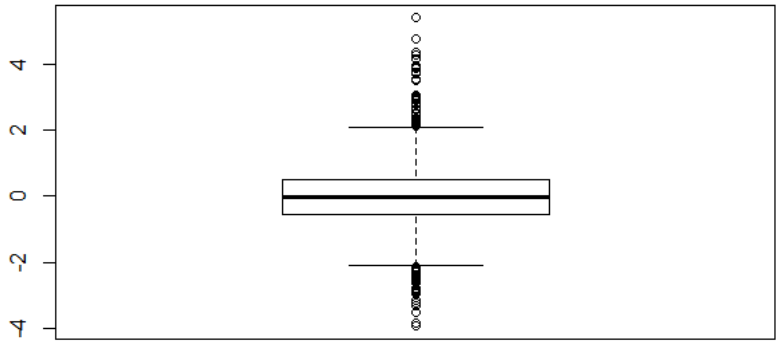
STEP 3　入出力関係：対数総負債=α_2+β_{13}対数総資産+β_{23}対数総資産+ε_2の外れ値



標準化残差2.0以上の3社
HIRABO, Nichibei Fuji Cycle, SEA-COM

- 標準化残差-3.0以下の32社
- SAKATA SEED, TEIKOKU OIL
  Kanto Natural Gas Development
  KITA KYUSHU COCA-COLA BOTTLING
  Shin Nippon Spinning, MIYUKI KEORI　ATSUGI
  NYLON INDUSTRIAL. SOTOH　Hokuriku Seiyaku.
  ASAHI DIAMOND INDUSTRIAL　SHIMA SEIKI MFG., NISSEI
  INDUSTRIES　SHINKAWA, TSUBAKI NAKASHIMA
  MABUCHI MOTOR., ICOM, KEYENCE　Japan
  Digital Laboratory, FANUC, NIPPON CONLUX
  FUTABA, Sony Music Entertainment(Japan), TENMA
  TOKYO STYLE, Fuji Kisen
  NAKANIHON THEATRICAL, KOEI
  TOKYOTOKEIBA, TOKAI KANKO
  Ga-jo-en Kanko, Sumisho Computer Systems　INES,
  SHINDEN

STEP 4　入出力関係：対数営業収入=α_3+β_{14}対数総資産+β_{24}対数総資産 +β_{34}対数総資産+ε_3の外れ値摘出



標準化残差3.0以上の18社
総合商社と水産関係
ITOCHU, Marubeni, TOMEN, Nichimen
KANEMATSU, CHUO GYORUI. MITSUI & CO.
TOHTO SUISAN, TSUKIJI UOICHIBA
OSAKA UOICHIBA, DAITO GYORUI,
SUMITOMO　Mitsubishi, SEIKA, Nissho
Iwai, TOKYO SANGYO　CHUBU SUISAN,
SHINKO GYORUI.

- 標準化残差-2.5以下の29社
- 鉄道業が抽出
- Chugai Mining, ORIX INTERIOR　KYOWA HAKKO KOGYO,
  Green Cross　INTERNATIONAL REAGENTS,
  ISEKI & CO. SANYO ELECTRIC,
  HANKYU REALTY HOKKAIDO SHINKO, Diamond City
  MITSUI REAL ESTATE SALES, TOBU RAILWAY
  Keihin Electric Express Railway
  Odakyu Electric Railway KEIO TEITO ELECTRIC RAILWAY
  Keisei Electric Railway, IZUKYU
  East Japan Railway, Kinki Nippon Railway
  HANSHIN ELECTRIC RAILWAY
  Nankai Electric Railway, Kobe Electric Railway
  Nagoya Railroad, Sanyo Electric Railway
  Nihonbashi Warehouse, WESCO
  Koshien Tochi Kigyo, DAI-ICHI HOTEL
  KYOTO HOTEL

The Scope of Quantitative Researches
Data Analysis II (1)

Principles of Quantitative Researches (1)
注）質問票の設計と教室内模擬調査
共分散構造モデリング

Hiroe TSUBAKI

2009/8/29

MBA-IB, GSBS

この年は、国際経営プロフェショナル専攻、経営システム科学専攻、企業科学専攻合同

Objectives
-Model-based Quantitative Survey-

- To understand
  - Definition of “Concept”
  - How to Quantify Concepts
    - Qualitative Approach
    - Model based Approach
  - Quality of Measurement
  - Relationship among Concepts
- by qualitative model-based thinking

Virtual Market Research Project
Group Work

- Organize Project Teams with 4 or 5 members (10/06: 1st class) at random
- ***Group-work 0***: Choose a product of which function or usage is fairly simple in each project (10/06)
- Clarify the usefulness (recognized quality) of the product from the viewpoint of the customer along the group-work 1~5 (10/06)
- Develop a measurement method of the level of the usefulness from the viewpoints of customer(10/13: 2nd class)
- Develop an evaluation method of association between the implemented function of the product and its CS (Customer Satisfaction) (10/20: 3rd class)
- Design a CS survey Excel sheet of several product profiles (by 11/10)
- Collect Data from persons of other project teams (11/10: 4th class)
- Analysis of Data (by 11/24)
- Presentation by each Group and Discussion on the above process (11/24: the last class)

Project Registration Sheet

- Project Team:
- Members and their e-mail address:
  - Leader of 1st class
  - Leader of 2nd class
  - Leader of 3rd class
  - Leader of 4th class
  - Leader of 5th class
- Product or Service :
- Basic Function of the Product:

Measurement Process of a Concept
Chapter 6, Bollen (1989)

- Give the meaning of the concept
- Identify the dimensions and latent variables or factors to represent the concept
- Form Measures
- Specify the relation between the measures and the latent variables

Step 1
Give the meaning of the concept

- Developing a theoretical definition
  - A theoretical definition explains in as simple and precise terms as possible the meaning of a concept
- Using not statistical methods but “terminology （用語学）” to develop it

Elements of Terminology

- Objects
- Concepts
- Terms
- Definitions

A Theoretical Definition of “Terrorism”

- The threat or use of violence for political purposes by individuals or groups, whether acting for, or in opposition to, established governmental authority, when such actions are intended to shock or intimidate a target group wider than the immediate victims
  - From the US Central Intelligence Agency’s 1981 report “Patterns of International Terrorism”

Ex.1 What is the intention of “terrorism”?

- Threat or use of violence
- For political purposes by individuals or groups
- Acting for or against established government authority
- Intended to shock or intimidate a target group wider than the immediate victims

Ex.2 Make a checklist to identify whether events below are terrorism or not.

- Hijacking planes and taking hostages
- Car bombs that kill civilians in London or Beirut
- Gangster-style killings
- Nazi’s slaughter of Jews during World War II etc.

Checklist

	Violence	political purposes	For or against government authority	Intended to shock or intimidate a target group wider than the immediate victims
○Hijacking Planes	Yes	Mostly	Mostly	Mostly
○Car Bombs	Yes	Mostly	Mostly	Mostly
×Gang-style Killing	Yes	No	No	No
△Nazi's Slaughter of Jews	Yes	Yes	Yes	?

Group-work 1

- What is the intention of usefulness of the product of your project?

Relationship between two Concepts

- **generic relation**
  - relation between two **concepts** where the **intension** of one of the concepts includes that of the other concept and at least one additional **delimiting characteristic.**
    - A generic relation exists between the **concepts** 'word' and 'pronoun', 'vehicle' and 'car', 'person' and 'child'.
- **delimiting characteristic**
  - **essential characteristic** used for distinguishing a **concept** from related concepts

Partitive relationships

- **Partitive relation (**part-whole relation)
  - relation between two **concepts** where one of the concepts constitutes the whole and the other concept a part of that whole
    - A partitive relation exists between the **concepts** 'week' and 'day', 'molecule' and 'atom'.

Structural Relationships

- **associative relation (**pragmatic relation)
  - relation between two **concepts** having a nonhierarchical thematic connection by virtue of experience
    - 'education' and 'teaching', 'baking' and 'oven'.
  - **sequential relation**
    - **associative relation** (3.2.23) based on spatial or temporal proximity
      - 'production' and 'consumption', etc.
  - ◎**causal relation**
    - **associative relation** (3.2.23) involving cause and itseffect
      - 'action' and 'reaction'

Group-work 2

- Create generic relationships of concepts in your product concept around “usefulness”

Group-work 3

- Make a set of concepts around your “usefulness”, clarify their characteristics and analyze their relationships using the concept diagram as shown in ISO 9000 if possible.

Not definition but measurement of a concept

To measure the level of “usefulness”

Step 2-1
Identify the dimensions of a concept

- Dimensions of a Concept
  - Distinct aspects of a concept
    - Components that cannot easily be subdivided into additional components
  - Dimensions of “Terrorism”?
    - Crosscutting
      - Whether an act is anti- or pro-government
      - Whether it is perpetrated by an individual or a group
    - →Four dimensions

Group-work 4

- Identify the dimensions of your concept as “usefulness”
  - Usefulness of a product
    - (effectiveness + safety + ・・・) ×
      (function A+ function B+・・・)

      ＝ effectiveness of function A
      + effectiveness of function B + …
      + safety of function A+ safety of function B + ・・・

Step 2-2
Identify latent variables to represent a concept

- Set one latent variable (or latent factor) per dimension
  - Terrorism
    - Factor 1: Antigovernment terrorism by groups
    - Factor 2: Antigovernment terrorism by an individual
    - Factor 3: Pro-government terrorism by groups
    - Factor 4: Pro-government terrorism by an individual

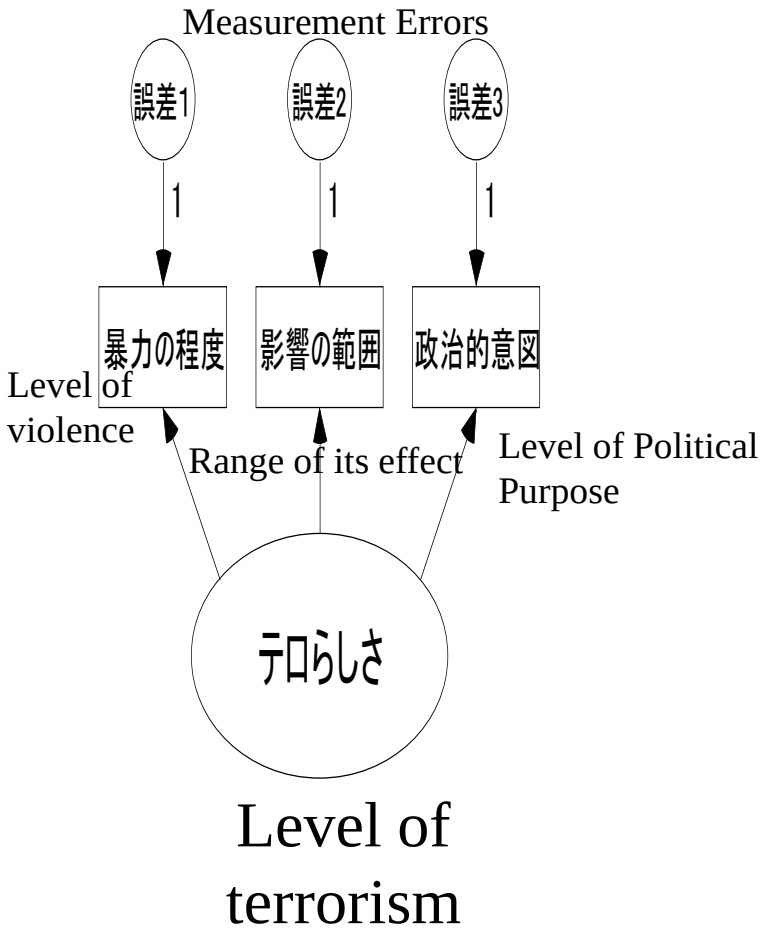
Step 3 Form measures
Operational definition of a factor

- The procedure to follow to form measures of the latent variables (factors) that represent a concept
  - To select observed variables that correspond to the meaning assigned to a concept
    - Responses to questionnaire items
      - Not checklists but questionnaire items asking level of the characteristics (Likert Scale)
        - Clearly Yes, Generally Yes, Maybe Yes, Neutral,
          Maybe No, Generally No, Clearly No
    - Quantitative measurements related to the characteristics

Step 4 (Next week)

Specify the relation between the measures and the latent variables

- Constructing the measurement model
  - A measurement model specifies a structural model connecting latent variables to one or more measures or observed variables



Ex.4 Construct an example of measures of the latent variable corresponding to antigovernment terrorism by groups

Measures	Yes	May be Yes	Un-known	Maybe No	No
Is it threat or use of violence?	x				
Is its purpose political?		x			
Is it an opposition to the governmental authority?					x
Is its intention to shock or intimidate a target group wider than the immediate victims?			x		
Is it done by groups?	x				

Group-work 5

- Construct an example of measures of the latent variable corresponding to some dimension of your usefulness or competitiveness

Presentation at 10/13

- Make presentation for 5 minutes by each group at the beginning of the next class to explain its tentative report on the group works, as
    - Process of defining concepts
    - How to derive dimensions
    - How to construct measures

標準データセットと
データ分析コンペティション

SSDSE2018の作成：筑波大学ビジネス科学⇒財務データ

Good Practiceの共有

品質管理学会地方研修データ⇒SSDSE（教育用標準データセット）

SSDSE： Standardized Statistical Data Set for Education （教育用標準データセット）

・都道府県・市区町村のすがた（社会・人口統計体系）からデータを抽出し、統計センターが2018年６月27日に公開（ https://www.nstac.go.jp/SSDSE/ ）

SSDSEの内容：地域プロファイル抽出にも役立つ

・縦に市区町村、横にデータ項目を並べた２次元の表形式データ
（1741市区町村 × 111項目、エクセルデータ 及び CSVデータで提供）

・1741市区町村（東京23区を含む）の内訳：791市、 744町、183村、23区
→　単純合計すると全国計に

・111項目は、複数の分野を１つのデータセットでカバー
（人口・世帯、自然環境、経済基盤、行政基盤、教育、文化・スポーツ、居住、健康・医療、福祉・社会保障）

・欠測値がない完備データセット

SSDSEの狙いと特徴: 自由課題へ進む前に自由度の高い規定課題を全国共有

・簡便性：容易にダウンロードでき、特別な前処理は不要（行・列の取捨選択のみ）

・親近性：学生や教員になじみがある「自分たちの地域を含むデータ」

・具体性：個別データについても意味が分かり議論できる

・多様性：様々な課題抽出、多様な分析が可能　→　自由度の高い「標準データ」

SSDSEに含まれる変数（2018年版）

【人口・世帯】
人口総数（合計、男、女）
日本人人口（合計、男、女）
15歳未満人口（合計、男、女）
15～64歳人口（合計、男、女）
65歳以上人口（合計、男、女）
75歳以上人口（合計、男、女）
外国人人口
出生数
死亡数
転入者数
転出者数
世帯数
一般世帯数
一般世帯人員数
核家族世帯数
単独世帯数
65歳以上の世帯員のいる核家族世帯数
高齢夫婦のみの世帯数
高齢単身世帯数（65歳以上）
婚姻件数
離婚件数

【自然環境】
総面積（北方地域・竹島を除く）
可住地面積

【経済基盤】
事業所総数
事業所数（産業大分類別（17））
第１次産業事業所数
第２次産業事業所数
第３次産業事業所数
従業者総数
従業者数（産業大分類別（17））
第１次産業従業者数
第２次産業従業者数
第３次産業従業者数

【行政基盤】
経常収支比率（市町村財政）
実質公債費比率（市町村財政）
歳入決算総額（市町村財政）
地方税（市町村財政）
歳出決算総額（市町村財政）
民生費（市町村財政）
土木費（市町村財政）
教育費（市町村財政）
災害復旧費（市町村財政）

【教育】
幼稚園数
幼稚園在園者数
小学校数
小学校教員数
小学校児童数
中学校数
中学校教員数
中学校生徒数
高等学校数
高等学校生徒数

【文化・スポーツ】
公民館数
図書館数

【居住】
総人口（非水洗化人口＋水洗化人口）
非水洗化人口
小売店数
飲食店数
大型小売店数

【健康・医療】
一般病院数
一般診療所数
歯科診療所数
医師数
歯科医師数
薬剤師数

【福祉・社会保障】
保育所等数
保育所等在所児数

統計データ分析コンペティションについて

SSDSEを用いた統計分析のアイデアと技術を競う論文投稿型コンペティション

→　Good Practice の共有化（論文審査後出版）

報道発表「児童・生徒等の統計リテラシー向上のための取組を実施します」
2018年６月26日、「キッズすたっと」開発の件と併せて総務省統計局で公表
（ http://www.stat.go.jp/info/guide/public/houdou/pdf/ho180626.pdf ）

※　以下は2018年の情報（2019年主催団体に統計数理研究所）

主催：総務省統計局、独立行政法人**統計センター、**一般財団法人**日本統計協会**

後援：国立研究開発法人**科学技術振興機構（JST）、**一般社団法人**日本統計学会、全国統計教育研究協議会、（2019年から全国高等学校校長会）**

主な日程：2018年６月26日　エントリー及び論文募集開始
８月10日　エントリー締切り
９月18日　論文提出締切り
10月18日　受賞論文の発表（統計の日）
11月19日　受賞者の表彰式（全国統計大会）

募集部門：高校生の部……高校、高専（1～3年次）の生徒
大学生・一般の部……短大、高専（4、5年次）、大学、大学院の学生、並びに一般の方

考察や結論を導く分析手順の必要十分な記述：教育成果様式の標準化

- **論文の構成**
  - 論文の構成については、原則として以下の構成に従ってください。
  - 第１章には、研究の目的と問題意識の背景を簡潔に記述してください。更に、提出された論文が参考にした先行研究があればその概要を記してください。
  - 第2章には、研究の方法（あてはめた統計モデル等）と手順を簡潔にまとめてください。
  - 第3章には、データセットからのデータの抽出、データセットへの変数の追加とその出典、分析に用いた変数に行った変換や加工などを記載してください。また、必要に応じてデータ分析に用いる変数の分布・要約統計量などについて、図表などを用いて分かりやすく示してください。
  - 第4章は、データ分析の結果等を可能な限り図表を交えて、分かりやすく記述してください。
  - 第5章では、得られたデータ分析の結果の解釈、または、分析自体の妥当性や限界などについても必要に応じて触れてください。その上で、結論を分析結果の独自性・新規性や社会に対する提言などの主張も含めて簡潔にまとめてください。
  - 最後に参考文献のリストも必要に応じて記載してください。
- **手順や分析結果などの記載の簡潔性**
  - ソフトウェア等を用いたデータ分析の出力を形式的に論文に全て貼り付けるのではなく、結論を導いた分析結果が第三者にも再現できることを意識して、必要十分な分析手順を記載してください。このテンプレートで用いているように、章を2.1節、 2.2節と見出しをつけて論理的に構成することも工夫してください。

2018年度高校生の部　受賞論文

受賞論文	受賞者	受賞論文の概要
【総務大臣賞】 本当に日本の医療は危機的状況にあるのか？ 雑誌統計2019/1月号掲載	広島大学附属高等学校 大段利々子	医療問題について、医師数、病院数と人口構成比や自治体の経済力などの関係性を分析した上で、高齢者の地域経済への貢献が重要と指摘
【優秀賞】 SSDSEデータを活用した全国学習状況調査結果との相関分析 雑誌統計2019/5月号掲載	和歌山県立田辺工業高等学校 宮本雨月、金山瑠依、門脇俊樹	小学生の学習への関心や取組について、家族構成や自治体の教育費等との関係を分析した結果、大家族世帯が学習状況に良い影響を与えると指摘
【日本統計協会賞】 交流人口増加による愛媛県の活性化 雑誌統計2019/2月号掲載 慶応大コンペでも受賞	愛媛県立松山南高等学校 白石大悟、高田蒼大、武田裕喜	人口減少による経済の縮小に対して、外国人旅行者の増加を目指し、温泉旅館の利用促進を提案。様々な公的統計を活用し、経済波及効果も推計
【特別賞】 機械学習による15歳未満人口の推定	渋谷教育学園幕張高等学校 伊藤寛子	機械学習により人口統計データを分析することを試み、その過程を詳細に記載

高校生の部　審査委員長の講評から（抜粋）

総務大臣賞「本当に日本の医療は危機的状況にあるのか？」

広島大学附属高等学校　　大段利々子さん

医療問題について、医師数、病院数と人口構成比や自治体経済力などの関係性を分析し、高齢者の地域経済への貢献が重要と指摘している。構成も論理的で読みやすい論文である。また、自治体人口当たり医師数について、都道府県間変動と県内の市町村間変動を比較した視点は、高く評価できる。ただし、一部の県内変動を大きく見せているのが、人口が小さな地域に大学附属病院が設置されているという外れ値現象に起因していることが、大学で学ぶ「箱ひげ図」で書き直すと明らかになる。さらに、相関の有意性の主張も、大学の学習で習得することができる。

日本統計協会賞「交流人口増加による愛媛県の活性化」

愛媛県立松山南高等学校　　白石大悟、高田蒼大、武田裕喜さん

地方活性化のために交流人口概念に注目して、 SSDSE以外の様々な公的統計を活用し、外国人観光客誘致に関する分析と対策に至るまでを考察した意欲的な論文である。愛媛県の訪日観光客について松山空港との関連性を仮説として提示したことは興味深い。ただし、筆者たちが着目した観光庁統計によれば、愛媛県の訪日観光客の入国空港は、関西国際空港、成田国際空港、高松空港の比率が高く、四国における高松空港の意味など、今後も考察を深めることが期待できる。

※「コンペティション」のHP 及び月刊誌「統計」掲載記事（2019年1月号 2月号）から抜粋

2018年度大学生・一般の部　受賞論文

受賞論文	受賞者	受賞論文の概要
【総務大臣賞】 地方創生における三つの「鍵」 雑誌統計3月号掲載	早稲田大学人間科学部人間環境科学科 平原幸輝	地方創生のコアである人口増減について、「大学進学率」、「人口当たり医師数」、「労働力人口率」の３つの指標が強い影響力を持っていると指摘。これらの指標に基づき、各市町村を類型化
【優秀賞】 人口規模によって異なる保育所数・保育所在所児数・定員充足率の関係	早稲田大学大学院教育学研究科 小野島昂洋	保育所数、利用者数、定員充足率は市町村の規模によって、その関係性に差異があることを確認した上で、政策立案においてはこの差異を考慮する必要があると指摘
【日本統計協会賞】 地方創生に向けた東京一極集中是正のための定量的都市圏選定指標の提案 雑誌統計4月号pp.55-60	慶應義塾大学リーディングプログラム 池田泰成、柴辻優樹、鶏内朋也、 石川貴啓、佐野岳史	首都機能を他の都市へ移転することを想定し、地理、経済、生活に係る指標を使って、移転にふさわしい都市を具体的に選定
【特別賞】 日本の全市町村における人口の自然増減の分布と説明要因	国際基督教大学社会科学研究所　小野恵子 Code for Nagoya, OSGeo財団日本支部 宮内はじめ 名古屋工業大学大学院工学研究科 白松　俊 名古屋大学大学院工学研究科　河口信夫 パーソルキャリア株式会社　五十嵐康伸	人口の自然減少について、地理的・時間的な分布を明らかにするとともに、高齢者率、若年女性比、女性・子ども比、都市圏かどうかが影響していると指摘

2019年度　高校生の部　受賞論文

（https://www.nstac.go.jp/statcompe/award.html）

受賞者	受賞論文（タイトル及び概要）
【総務大臣賞】 竹内遥・江本もえ・木下舞・永井あゆる （お茶の水女子大学附属高等学校） 雑誌統計2020/01掲載確定	ワンオペ育児から見る離婚 離婚の要因を探るため、様々な仮説の下、総人口の影響を除いた偏相関係数を用いた相関分析を行った。その結果、離婚要因の一つが、家庭内で女性のみが家事や子育てを行うワンオペ育児にあることを導いた。その上で、ワンオペ育児を防ぐために、男性の育休取得数を増やすことなどを提案している。
【優秀賞】 渡邉璃里香・吉田美咲 （愛媛県立松山南高等学校） 雑誌統計2020/02掲載確定	南海トラフ地震に備えて　～指定避難所に3人に1人が避難できず、災害時の医療体制は本当に十分か？～ 南海トラフ地震に備えるために、通学している高校から半径3km以内のエリアについて、GISを用い地図上に指定避難所や診療所をプロットすることにより、避難所の分布に空白地帯があり診療所に偏りがあることを指摘した。さらに、幼稚園等の新たな避難所、災害時の医療体制の充実などを提案している。
【統計数理賞】 猪狩信人 （福島工業高等専門学校） 雑誌統計2020/07掲載予定	過疎地域の現状分析と発展に重要な視点 過疎地域の発展に重要な視点を探るため、市区町村を過疎地域とそれ以外に分類し、SSDSEから人口、教育、産業などのデータについて、分布の比較や相関分析を行った。その結果、一次産業の活性化が過疎地域の発展につながることを指摘するとともに、行政機関が資金援助を行うことの必要性を述べている。
【統計活用奨励賞】 大段利々子 （広島大学附属高等学校） 雑誌統計2020/05掲載予定	日本で暮らす外国人の動向から見た多民族化 多民族化が人口問題解決の鍵になるという仮説の下、SSDSEから人口、地方経済関連データを抽出して相関分析を行った。その結果、外国人比率の高い地域が都市部と地場産業を有する地方とに二極化していることなどを示し、新たな産業の展開が外国人の増加、地方創生につながる可能性を指摘している。

2019年度　大学生・一般の部　受賞論文

（https://www.nstac.go.jp/statcompe/award.html）

受賞者	受賞論文（タイトル及び概要）
【総務大臣賞】 張瀚天・白鳥友風 （筑波大学大学院システム情報工学研究科） 雑誌統計2020/03掲載予定	地方創生目標指標に関する変化要因ネットワークの推定とそれに基づく地域間連携策の提案 地方創生の設定目標について、目標指標に影響を与える変数の相互関係を明らかにするため、共分散行列を用いて要因ネットワーク図を作成し、lasso法による分析を行った。その結果、地域の稼ぐ力が地方創生に重要であることを指摘するとともに、地理的な制約にとらわれない地域間連携を提案している。
【優秀賞】 竹内太郎 （大阪大学医学部） 雑誌統計2020/04掲載予定	我が国における人口増減の決定要因 都道府県の人口増減について決定要因を探るため、教育、健康・医療分野の指標を説明変数として回帰分析を行った。その結果、高等学校卒業者の進学率、一般診療所数、医師数の増加などが人口増加に影響しているとの結果を得ており、人口減少を克服するための政策立案の基礎資料として提案している。
【統計数理賞】 松本洋輔 （一橋大学経済学部） 雑誌統計2020/08掲載予定	マルチレベル分析を用いた市町村大学等進学率の決定要因分析 大学等進学率の地域格差の要因を探るため、都道府県レベルと市区町村レベルの複数の説明変数を同時に扱うマルチレベル分析を行った。結果として、都道府県レベルでは東京または京都までの距離と大学収容率、市区町村レベルでは課税対象所得、知識集約型産業従事者率等が、大学等進学率に影響していると指摘している。
【統計活用奨励賞】 村松波・熊野翔・川田瑛貴 （武蔵野大学工学部） 雑誌統計2020/06掲載予定	市区町村別でみる合計特殊出生率推移の特徴分析 少子化問題の特徴を明らかにするため、子ども女性比を用いて市区町村別に合計特殊出生率（TFR）の推定を行い、2015年までの増減について検証を行った。結果として、東京都区部では港区等の増分が大きいことと、TFRの増分は納税者一人あたり所得の増分との相関が高いことを指摘している。

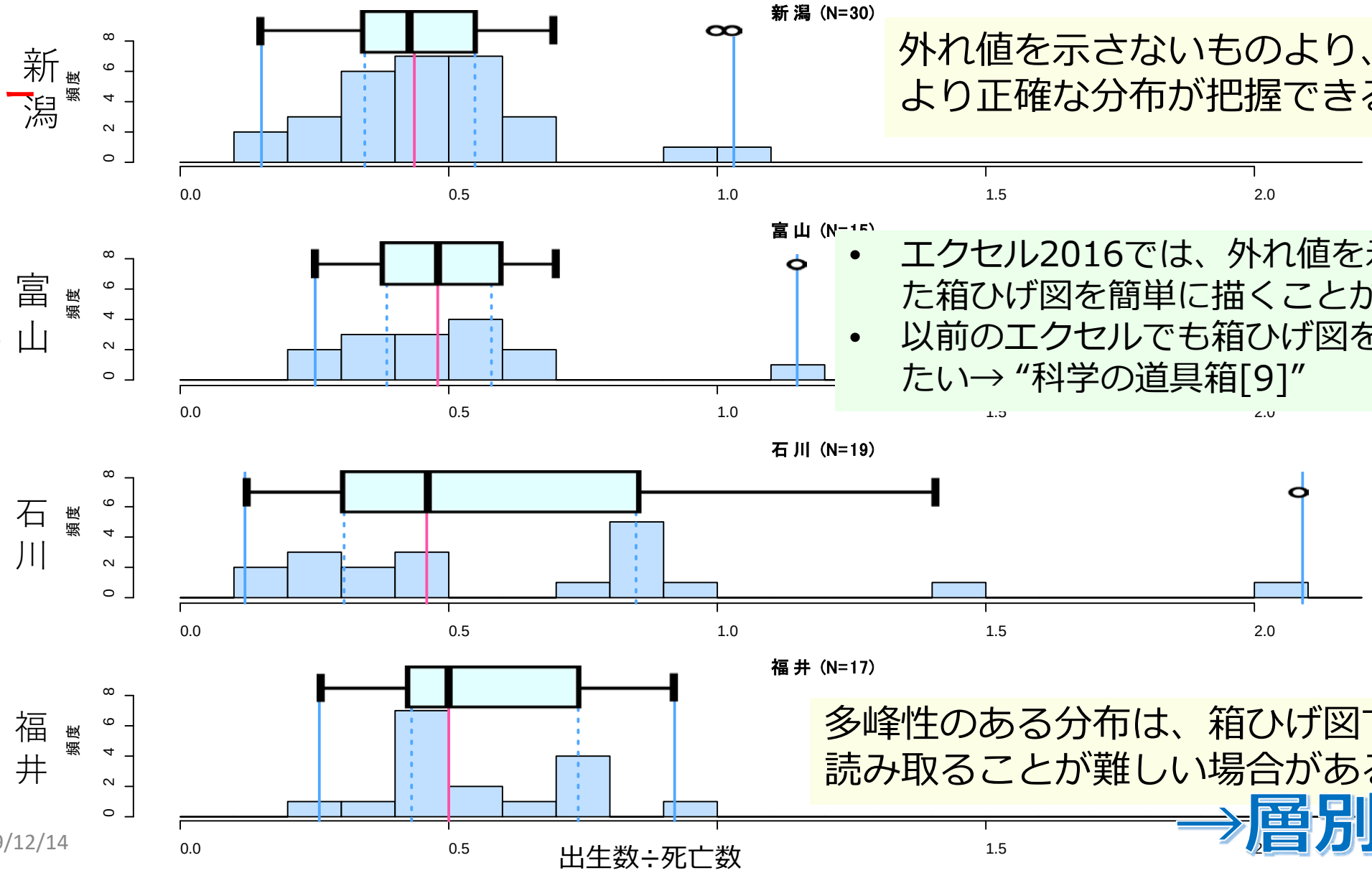
2019年度　特別賞 受賞論文

（https://www.nstac.go.jp/statcompe/award.html）

受賞論文	受賞者	受賞論文の概要
【高校生の部】 **香川県の交通事故発生の要因を交通違反件数を基に分析する**	香川県立観音寺第一高等学校 **宇川 昇吾、宮本 紫苑 山地 悠介**	香川県の交通事故発生件数が多いという問題意識の下、交通事故発生要因を解析するため、交通違反件数を説明変数とする重回帰分析を行った。その結果、香川県と人口規模が近い和歌山県との比較を通し、香川県では一時停止違反の多さが交通事故の発生に深く関わっていることを指摘している。
旅館及びホテルにおける日本人・外国人宿泊客の都道府県別増減から考える旅館の復活　―岡山県湯原温泉の視点からインバウンド需要を旅館に取り込む方策―	岡山県立岡山操山高等学校 **池田 雅子**	生まれ育った温泉地の復興に向けて、インバウンド消費が重要であるという仮説の下、GISソフトを用いた分析や相関分析などを行った。ビジネスホテルと外国人観光客数には正の相関が観察されることから、設備投資を行い、和風のビジネスホテルにするなど、地方の旅館再生モデルを提案している。
【大学生・一般の部】 **潜在患者数に対する医師偏在の可視化**	東北大学大学院文学研究科、株式会社社会情報サービス **眞田 英毅、三浦 萌実**	医療需要を踏まえた医師偏在の実態を把握するため、潜在的医療需要として入院患者数を二次医療圏別に算出した上でジニ係数を用いて検証を行い、都道府県比較を行った。結果として、医師偏在は、おもに西日本で発生していることを指摘し、都道府県内の医師の割振り配置について提案を行っている。
外国人人口と市区町村の特性との関係性	関西学院大学経済学部 **西尾 春香**	外国人人口と自治体の特性との関係について分析するため、市区町村別の統計指標を用いた重回帰分析を行い、人口密度や製造業等との相関が高いという結果を得た。また、決定木分析、ランダムフォレスト等の追加分析を行い、都市化、働き手不足、特定産業等が外国人人口に影響していることを指摘している。
「広域連携の政策検証」 ―空間計量経済学的手法による実証分析―	早稲田大学政治経済学部商学部、社会科学部 **原 康熙、福田 和生 柳田 はづき**	地方自治体の広域連携の効果について検証するため、Moran統計量による空間相関分析を行った。その結果、ごみ処理費用については広域連携の効果が確認され、待機児童についても共通の政策目標が形成された自治体の広域連携においては待機児童数が減少し、行政の効率化を促す可能性を指摘している。
地方創生実現のロジック ―地域経済活性化のメカニズムを解明する―	早稲田大学大学院人間科学研究科 **平原 幸輝**	地域経済活性化のメカニズムを解明するため、市区町村別の付加価値額を事業従事者数等から重回帰分析により推計し、人口と付加価値額への影響をパス解析により分析した。結果として、労働・医療・福祉・教育といった社会環境の充実が、人口増加と経済活性化を引き起こし地方創生につながると指摘している。

ヒストグラムと箱ひげ図の比較(出生数/死亡数)：外れ値

**統計センター
山下雅代
研究員が
開発中の
教材.
雑誌統計
2019/6月号**



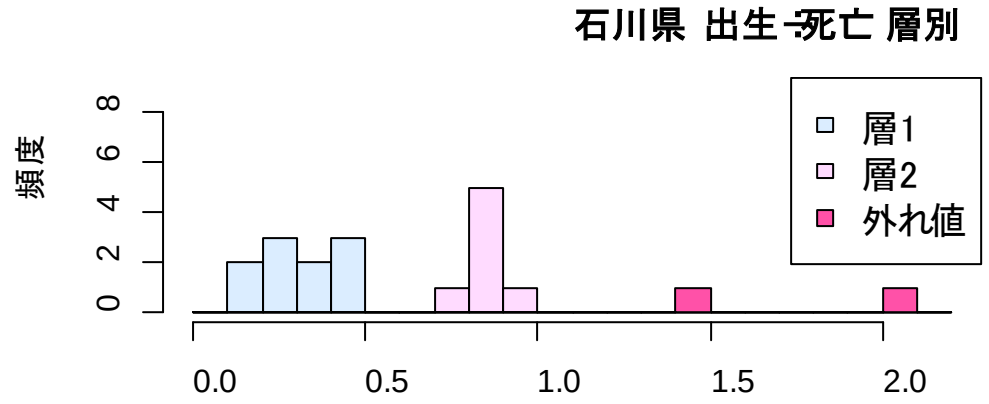
外れ値を示さないものより、
より正確な分布が把握できる

- エクセル2016では、外れ値を示した箱ひげ図を簡単に描くことが可能
- 以前のエクセルでも箱ひげ図を描きたい→“科学の道具箱[9]”

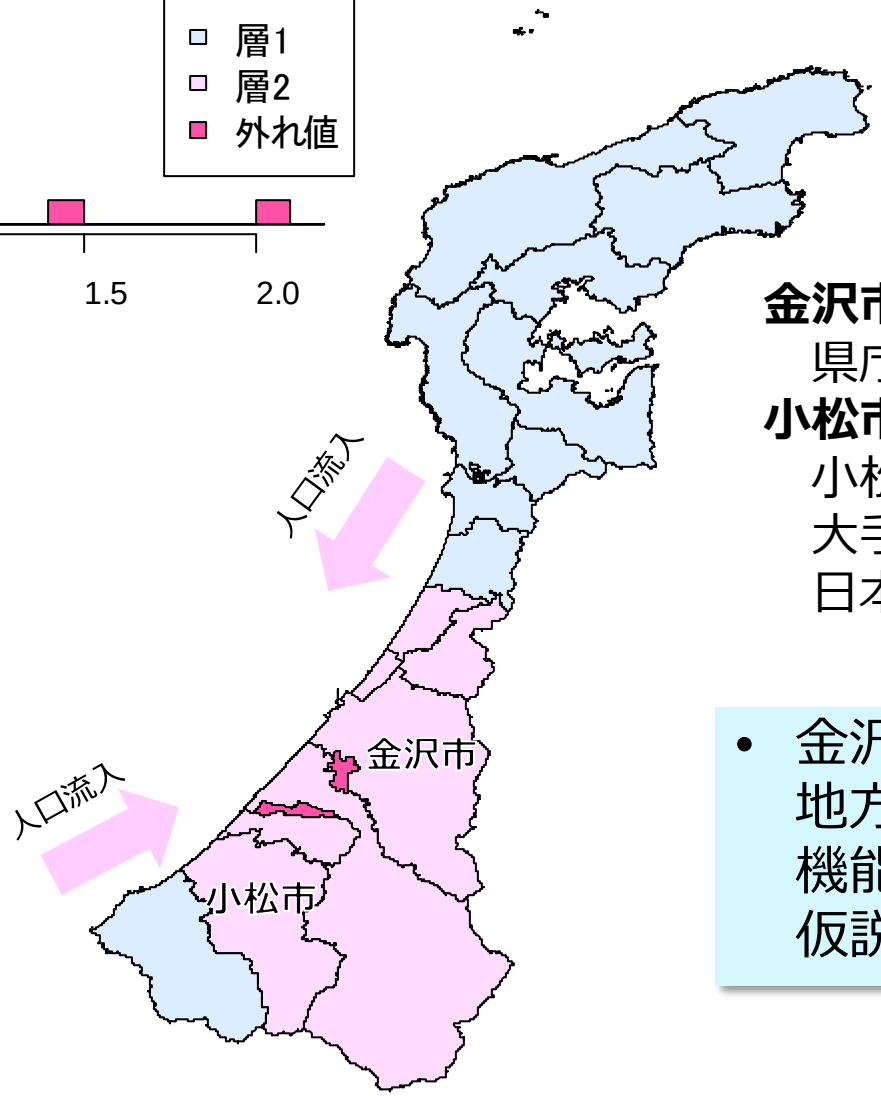
多峰性のある分布は、箱ひげ図で
読み取ることが難しい場合がある

⇒層別

石川県のデータの層別：3分類：統計センター山下雅代研究員作成



雑誌統計連載掲載予定教材



金沢市：
県庁所在地
小松市：
小松製作所の本拠地
大手企業の事業所が集まる
日本有数の企業城下町

- 金沢市と小松市の2市が地方の中核都市として、機能しているという仮説が立てられる

公的統計利活用のための人財育成基盤の3本の矢：今後の取組みの展開

SSDSEを中心とした統計教育支援の取組み

SSDSEの整備

・2019年版を６月頃に公開予定

・都道府県データや時系列データの追加も

ニーズを踏まえた
見直し・拡張

SSDSEを使った教材開発

日本統計協会の月刊誌「統計」へ連載

・**中等教育用教材：2019年4月（pp.49-54）～**

「授業に使えるSSDSEの統計教材（中学・高校編）」

・**大学用教材：2019年7月～**「データサイエンス入門」

統計データ分析コンペティションの継続

・2019年度に第２回コンペティションを開催

・**統計数理研究所** 共催団体へ，様々な連携模索

・SSDSEは2018年版、2019年版いずれも利用可

３つの相乗効果を期待

統計的機械学習への途と
これからのデータ分析教育

技術進歩の主要原理の理解

機械学習の幾つかの主要原理
なぜ機械はデータで賢くなるのかを認識すべき

○任意関数自動近似技法

- **多項式型回帰（応答局面法）➡ニューラルネットワーク系統計的機械学習**
- Friedmanの**射影追跡回帰（交互作用消去＋GAM）**＋不連続分布と連続分布同時近似可能な非線形関数
  - ボルツマンマシン：温度パラメータが高ければステップ関数（異質性自動分類）も近似可能
- **変数間交互作用の消去**と**非線形主効果（層別）**の合わせ技で任意関数近似：**$4XY = (X+Y)^2 - (X-Y)^2$**
- **古典的データ分析戦略の進化をニューラルネット系AIは概ね反映：ただしブラックボックス化**
  - **特性値を上手に選択**せよ：**因子モデル，PLS➡合成指標の最適探索**
  - **層別**を上手く行え：**潜在クラスモデル➡層別境界の自動探索**

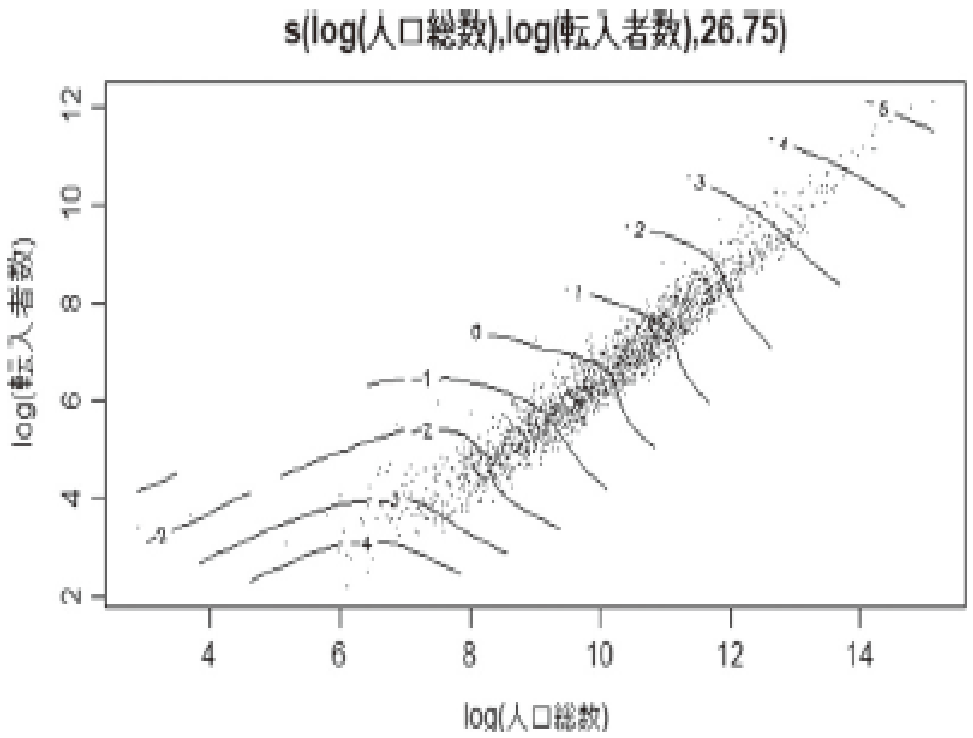
○近似関数の単純化（パラメータ節約）

- **高次因子モデル➡深層学習：**階層化による単純構造での近似
  - 合成指標による予測方式自体の階層的クラスター表現：実はパラメータ数節約している（そうは思われていない）

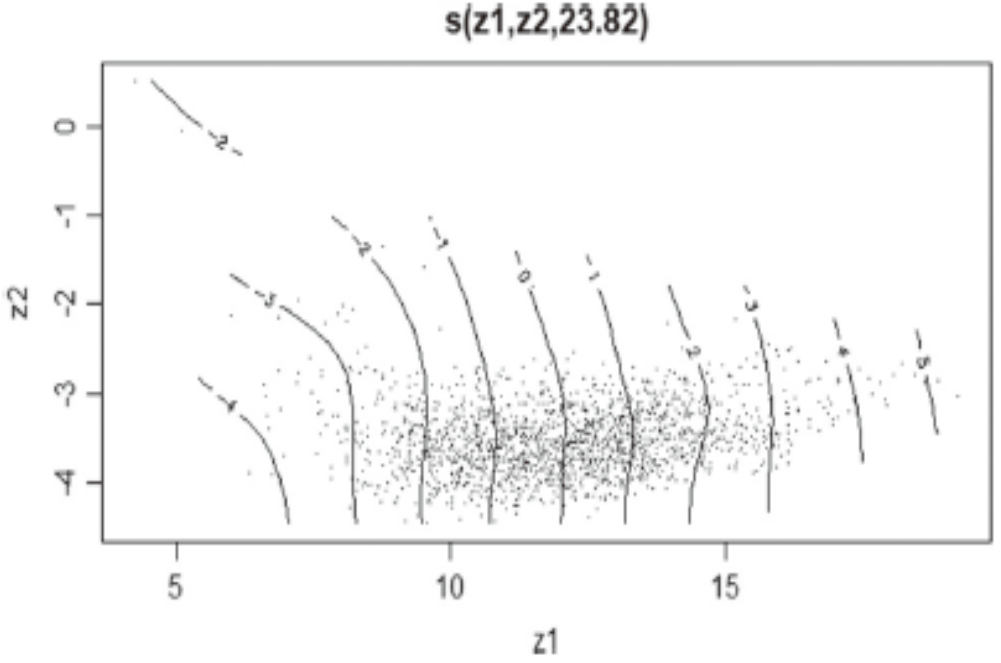
SSDSE による射影追跡回帰
婚姻件数の予測

$$f(x_1,\ldots,x_p)=\sum_{i=1}^{q}\beta_i g_i\left(\sum_{j=1}^{p}w_{ij}x_j\right)$$

2次元単純平滑化



射影追跡回帰：特徴量



$z_1=0.847\log(\text{人口総数})+0.532\log(\text{転入者数})$

$z_2=0.71\log(\text{転入者数}/\text{人口総数})$

正則化あるいはベイズモデル戦略

〇モデル選択➡
正則化＝
ベイズモデル事前分布の
最適化

- **許容性のあるベイズ決定の中で**
  パフォーマンス最適化

〇逆問題（制御問題）の
正則化への注意

- 非予測トレーニングデータの
  平均・分散を予測に用いるのか？
  - 田口T法，ナイーブベイズを見直すべき

予測値を束ねる戦略

○複数単純**予測値の情報量按分最適結合**

- 個人の知より集団の知
- **Bartlettの因子得点推定，田口玄一(品質工学)のT法が基本原理**
  - ➡アンサンブル学習, **Random Forest**

○予測誤差を改善する予測ロジックの追加戦術

- **回帰分析改善のPDCA**サイクルを回している
  - **回帰診断や**田口玄一の実験的回帰分析➡**Boosting**

（独）統計センター時計技術研究課
家計調査消費項目600分類への活用想定
機械学習研究

例：単純な 産業21分類教師付き学習

2値特徴量：(財務変数＋層別基準)を
そろえて1996年上場企業産業分類

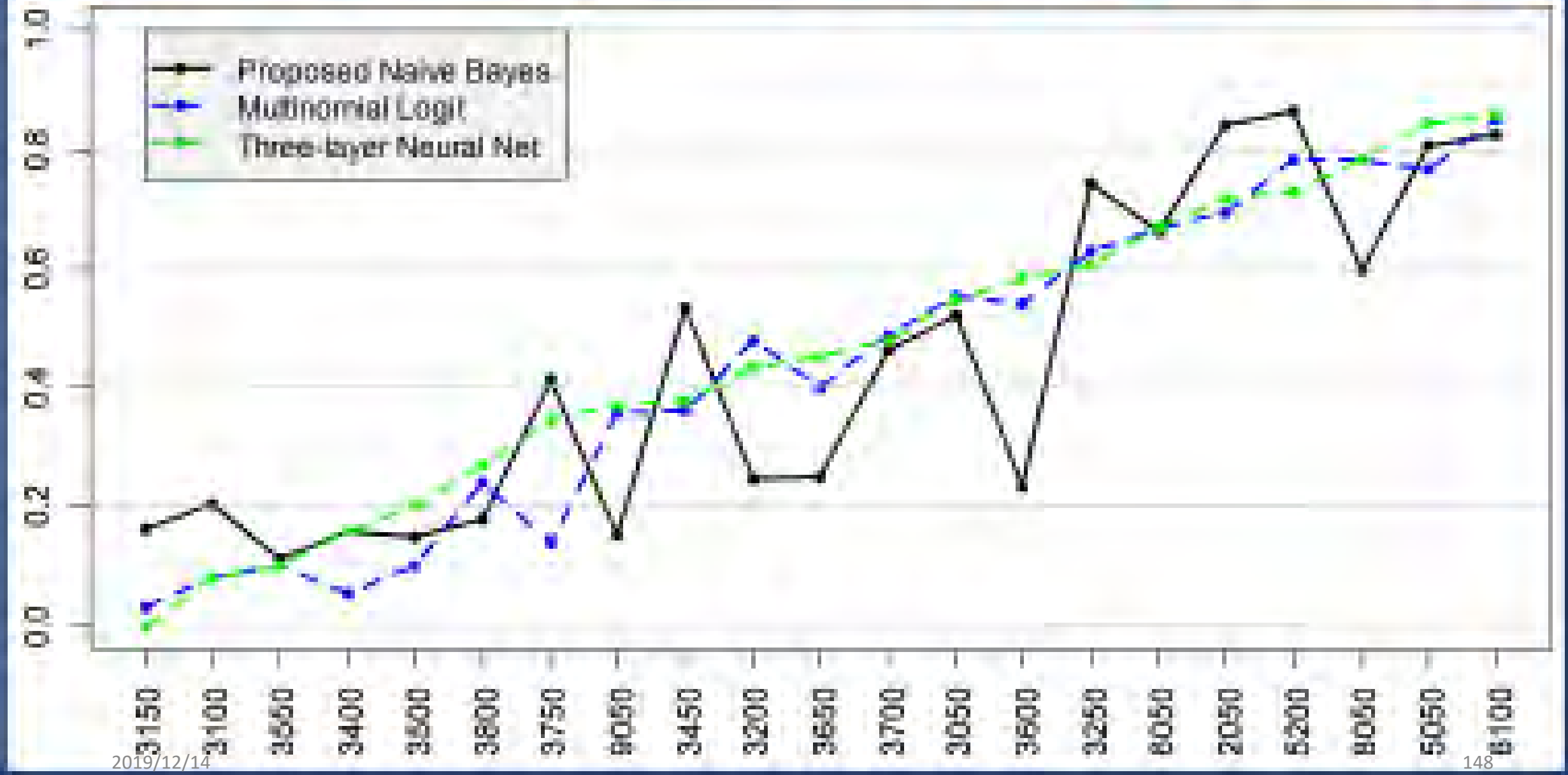
仮説：産業分類が財務情報に影響
逆問題：T法を離散データに拡張
　ミニマックス的・ナイーブベイズ

単純な順問題型予測
　多項ロジットモデル
　3層ニューラルネット

Table 1 Industry classes for the dataset classification

Industry Class	Securities Classification Code	Number of Companies
Construction	2050	177
Foods	3050	119
Textiles and Apparels	3100	98
Pulp and Paper	3150	31
Chemicals	3200	157
Pharmaceutical	3250	43
Glass and Ceramic	3400	57
Iron and Steel	3450	58
Non-ferrous Metals	3500	40
Metal Products	3550	70
Machinery	3600	192
Electric Appliances	3650	203
Transportation Equipment	3700	103
Precision Instruments	3750	29
Other Products	3800	67
Land Transportation	5050	51
Warehousing and Harbor Transportation	5200	37
Whole Sale Trade	6050	181
Retail Sale Trade	6100	149
Services	8050	37

Correct classification rates of the classifiers



計算機環境

☆パフォーマンス
最適予測探索に
資する計算機パワー

モデル選択➡価値関数
クロスバリデーション
評価などの
自動最適化

その他のアイデア

超高次元内での分類・
回帰の**効率化**

- カーネルトリック
（Support Vector Machine）

階層性を意識したデータサイエンス教育

初級者用

- 入力制約が少ない，
- 出力システム理解可
- 中程度の予測精度
  - **Classification and Regression Tree**：
    - 自動層別技法
    - 品質管理の基本

中級者用

- 入力に加工を行い，
- 出力システム理解可
- 中程度の予測精度
  - 交互作用や非線形性を考慮
    - セミパラメトリック回帰モデル
- 散布図に線を引く
- 通常の回帰モデル
  - 散布図に式を適合
  - **射影追跡回帰**

上級者用

- メカニズム解釈を含むモデルで高精度予測
  - 技術者が統計家との協働で作る非線形パラメトリック予測モデル
    - データ同化
- 一般化可能
- 技術知識の獲得

ベンチマーク用ブラックボックス

- 入力制約が少ない，システムは分からないが，ベストパフォーマンス
  - Random Forest
  - 深層学習
    - ブラックボックス
    - チューニング大変

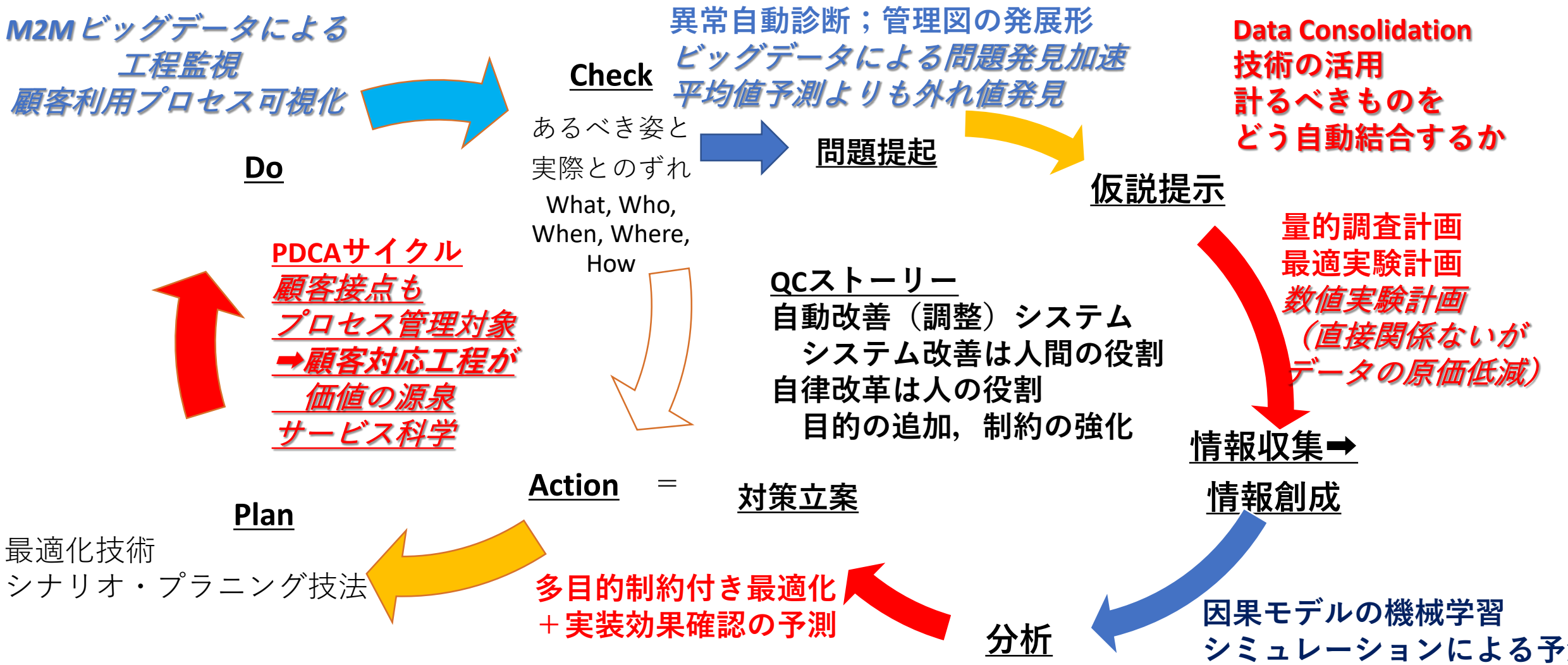
おわりに：
これからも変わらない
知の獲得プロセス教育を

データ分析教育の本質は認識とデザインのプロセス教育

分析技術はプロセスの特定のフェイズに適切に埋め込む

ユーザー教育はプロセスと技術・その使用上の注意を体験教育し経験を共有

デミング・石川モデルへのAIの埋め込み
挑戦：デミング・石川モデルのSHINKAの方向性



科学技術振興機構 平成23年度 モデリング分科会（2010）の活動 CRDS-FY2010-XR-20,https://www.jst.go.jp/crds/pdf/2010/CRDS-FY2010-XR-20.pdf

自身の関わった統計教育文献

- 土橋俊人、高須久、椿広計(1985)どう解析するかこのデータどう収集するかこの言語情報、品質、Vol.15(3),pp.56-66.
- 椿広計(1994)回帰分析から因果分析へ　計測データの分析で学んだこと(2), 要因分析と予測のための統計的アプローチ7、標準化と品質管理、Vol.47(5),pp 111-116
- 椿広計(1999)データサイエンスの社会人教育（特集データサイエンス第1部データサイエンス登場、KEIO SFC review, Vol.3(1), pp.38-43.
- 椿広計(2004)IT時代の統計教育：統計科学の目的と理解に向けて＜特集　IT時代の品質技術と品質管理教育＞, 品質、Vol.34(1), pp.48-56.
- 椿広計(2006)実学としての統計科学（特集・今に生きる「実学」）、三田評論、No. 1087, pp.22-30.
- 椿広計(2006)ビジネスへの統計モデルアプローチ、朝倉書店
- 椿広計(2009)経営プロフェッショナル教育の質保証：わが国ビジネススクールの試み＜特集　実務に役立つSQCの再普及＞、品質、Vol. 39(1), pp.18-24
- 渡辺美智子、椿広計編(2012)問題解決学としての統計学ー全ての人に統計リテラシーを、日科技連出版
- 椿広計(2016)　モデリングとその教育について, 科学教育研究, V0l.40(2), pp.119-126.
- 総務省政策統括官編(2017)**大学での学びにつながる統計で身近な現象や社会の課題を探究するスタディガイド高校からの統計・データサイエンス活用～発展編～**統計的思考力を身につけよう！日本統計協会
- 椿広計(2018) 小学校・中学校における算数・数学教育の中に如何にして統計的考え方を導入すべきか?,統計数理、Vol.66(1), pp.4-14.
- 椿広計(2019)データ駆動型社会の人と品と質とのマネジメント、応用統計学、Vol.47(2/3),pp.89-98.
- 山下雅代、椿広計、飯島信也(2019) 教育用標準データセット(SSDSE)による探究型統計教育の促進 : 総務省統計コミュニティの試み (特集 数学科におけるデータサイエンス(1)),日本数学教育学会誌、Vol.101(3), pp.40-47.
- 椿広計：データサイエンス入門：雑誌統計（日本統計協会）2019年7月号から連載

ご清聴ありがとうございました

データ・サイエンス時代でも人間はより賢くなるものと信じます
そのためにも良いデータ分析事例を社会共有することが大切です