

統計・数理モデリング 教育の目指すところ

椿 広計

独立行政法人 統計センター

内容：初中等統計的方法教育を対象として ジグザクした考察を行う

- はじめに
- 数理科学的方法の機能を使いこなすこと
- 不確かな現象のモデリング
- 無作為抽出標本と統計的推測
- 統計モデル当てはめプロセスと仮説形成
- 初等的な統計的機械学習初等教育の可能性
- 効率最適化に関する推論
- おわりに

はじめに

統計教育の重要性と設計

標準データセットによるモデリングの知の共有

統計教育の重要性と設計

- データの価値 = データから派生する知識価値（ニーズ依存）
ーデータとその処理の価格（負の経済価値）
 - コスト低減⇒データ駆動サービス産業革命
- 効率的・効果的知識価値創成の横断科学～計測と統計
- 育成すべき人材像と教育のデザイン
 - 生徒の活動，教員の活動，教程，教材，教育方法
 - 最終目標活動：完全創造的アクティブラーニング～自由課題
 - 目的を持った創造的情報収集と分析
 - 「科学の文法」教育：「最重要」：今晚の話題
 - 前段としての共通課題とその教材：
 - 自由度の高い「標準データ」とその自律的かつ多様な分析の生徒間での共有
 - 標準データセットの要求品質
 - 親近性「自分たちのデータ」：学生や教員になじみがあること
 - 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：様々な課題抽出が可能

2018/06/27 統計局・統計センターSSDSE Standardized Statistical Data Set for Education

<https://www.nstac.go.jp/SSDSE/>

• 都道府県・市町村のすがた（社会・人口統計体系）

- 全市区町村（東京23区を含む）：1741自治体
 - 791市, 744町, 183村, 23区
 - 生徒誰もがどこかに所属。有名な市町村も多い。
- 111変量抽出：完備データセットに加工
 - 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
- 高校**探究科目**, 大学データ分析教育, **Good Practice 共有化**への活用
 - 文科省の日本食品標準成分データベース（2198×63）

• 教育用標準データの量的階層性

- 都道府県データ・教室データ取得（50件規模）：小中学校
- 市町村データ, 上場企業データ（1500～2500件規模）：高校・大学
- 定型的ビッグデータ（数万件規模）：匿名ミクロデータ：大学院
- 画像・音声などのビッグデータ

標準データセットによる Good Modelling Practicesの知の全国共有

• 統計データ分析コンペティション

- 総務省，独立行政法人統計センター，一般社団法人日本統計協会共催
- 後援（予定）：国立研究開発法人科学技術振興機構，
一般社団法人日本統計学会，全国統計教育研究協議会
- 高校生の部（高等学校，高等専門学校1-3年次）
- 大学並びに一般の部
 - 個人またはグループ，指導教員の指導可
- エントリー期間：6月26日～8月10日
- 論文提出期間：6月26日～9月18日
 - 論文要旨＋論文（8ページ以内）を所定の様式で提出
- 受賞論文の決定・発表：10月18日（統計の日）
- 1月全国統計大会で表彰：総務大臣賞，優秀賞，日本統計協会賞
 - 優秀論文を雑誌「統計」に掲載することを検討中

数理科学的方法の 機能を使いこなすこと

2つの機能

モデリングとモデルの違い

数理科学的方法の2つの機能

- 数理科学・統計科学的方法の機能
 - 認識のための記述・表現機能
 - 数え上げ, 関数, 微分・積分
 - 設計のための評価・最適化機能
 - 積和計算, 比率計算, 方程式, 最大・最小
- 初中等数理科学教育
 - 数学に内在する基本機能を使いこなせる人材の育成
 - そのために必要な知識の伝達
 - 狭義のモデリング (モデルによる近似記述・表現)
 - 広義のモデリング (記述・表現の活用)
 - 統計的推測は広義モデリング

不確かな現象のモデリング

統計的方法の教育の目指すところ

よく分からないことの表現

統計的方法の教育の目指すところ

- 統計モデリングの要素
 - 不確かな現象の記述・表現
 - 統計モデルの当てはめと妥当性検証
 - 不確かな現象の評価・最適化
 - 統計的決定
- 不確かな現象の表現？
 - 系統的な変動の記述・表現 + 原因のよく分からない変動の記述・表現
 - **説明のつかない変動を確率分布で表現⇒伝統的統計モデル**
 - 記述統計による近似推測（シェーハートの原理）：ベイズモデル・変量模型
 - 突き止められる原因をコントロールし続けた残りの変動が偶然変動
 - **他にも様々な表現戦略**があることには留意
 - 複雑系・カオス（系統的表現），ファジイ（事象の混合：猫は半分死んでいる）

無作為抽出標本と統計的推測

データを確率的存在（確率変数）に変換
SSDSEの母集団情報として標本分布を体得

統計的推測の必要条件

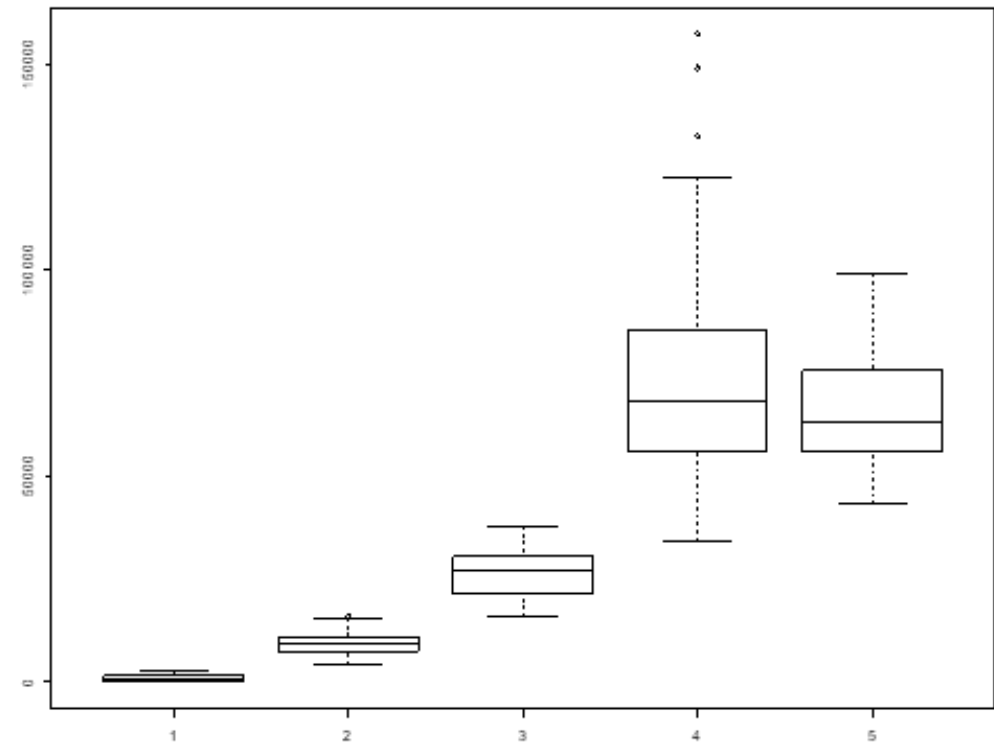
- 中学の統計的活動：標本抽出
 - ランダム・サンプリング：データを確率変数化する人為的技法
 - 統計量の標本分布：正解が分からない問題
 - 母集団情報既知のデータから無作為抽出
 - SSDSEを母集団情報として教育現場で利活用
- 要約統計量の標本分布
 - 統計的推測の背後に存在する分布
 - 母集団から繰り返し無作為標本抽出し，実際に標本分布を発生

SSDSE試作版から50自治体 無作為に40回抽出

表1 大きさ50の無作為標本の
人口に関する統計量

Data	最小値	25%点	中央値	平均値	75%点	最大値
標本 1	975	8109	19736	116139	53220	2691185
標本 2	832	7384	23686	50482	55257	519594
標本 3	18	8278	21333	51024	51848	483480
標本 4	1461	4910	15710	38434	40812	315814
標本 5	225	5249	15896	46228	42519	442913
母集団	0	8202	24622	73001	63431	3724644

最小値, 25%点, メディアン, 平均値,
75%点の標本サイズ50の標本における
標本分布 (大きさ40のシミュレーション)



シミュレーションによる統計的推測

統計量の標本分布 (大きさ40のシミュレーション)

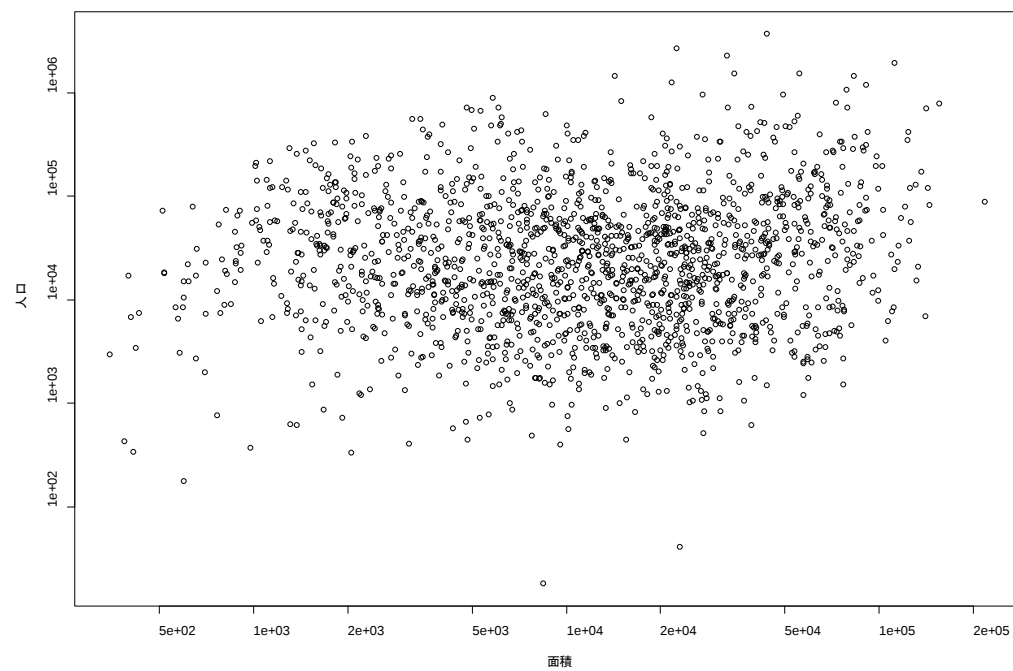
	標本分布を調べた統計量					
	最小値	25%点	中央値	平均値	75%点	最大値
最小値	0	4488	15886	34609	43365	174742
25%点	436	7600	21718	56343	55973	490229
中央値	853	9400	26964	68473	63400	704989
平均値	1091	9617	26491	74661	65780	1003210
75%点	1729	10944	30462	85614	75627	1082159
最大値	3092	16172	37966	157401	99119	3724844

統計量の標本分布

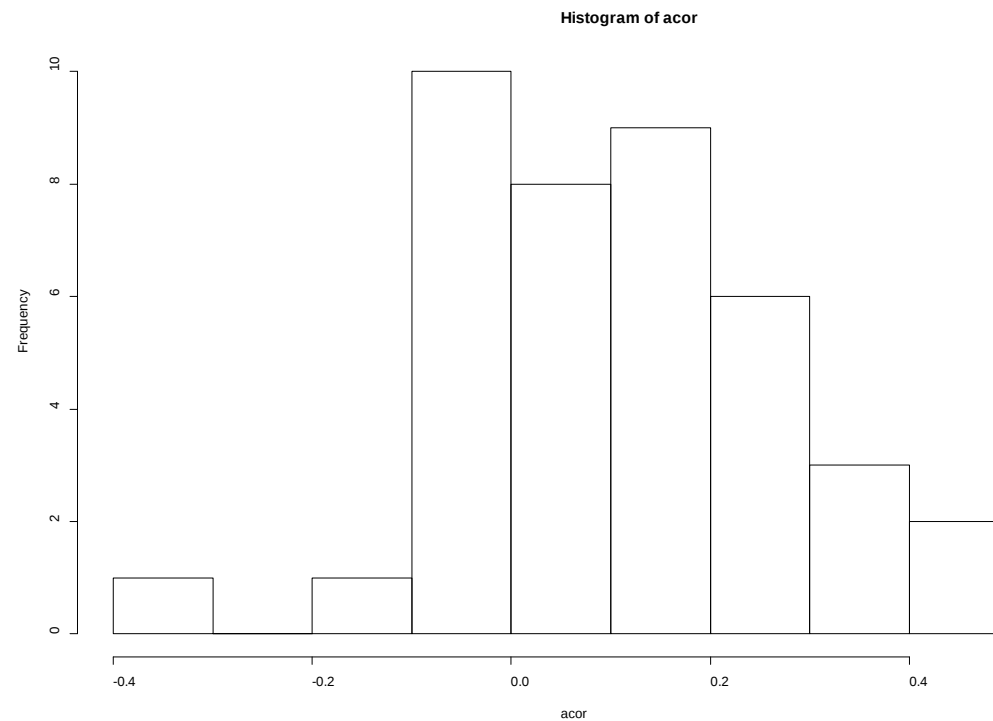
- 25%点から75%点までの区間
- 通常は50%信頼区間と考える
- 母集団の最大値
 - 標本分布の最大値より大きい
- 2標本検定への適用
 - 東日本と西日本から大きさ一定の標本を抽出
 - 平均値の差の統計量の標本分布
- 正規分布を想定した検定や信頼区間
- 二項分布の正規近似による比率の検定
 - シミュレーションを行わないでも理論的に統計的推測ができる方法
 - 簡便近似

厳密な分布は難しくても

自治体面積と面積の相関 母集団の散布図



無作為標本の大きさ50で相関係数を 40回計算した場合の標本分布



統計モデル当てはめの プロセスと仮説形成

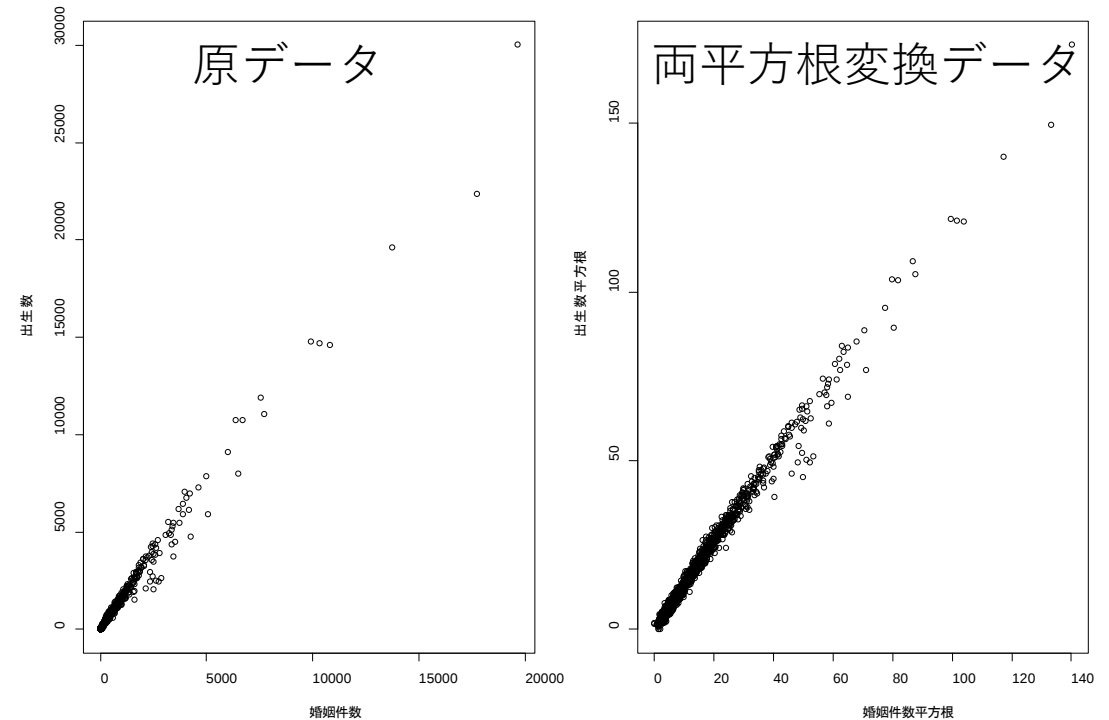
無作為化の行われていないデータへの統計モデル当てはめ
仮説成長型（機械学習型）接近とシステム科学的接近

一見推測統計風の記述統計

- 与えられたデータへの統計モデル当てはめ
- 残差分析で系統変動が残らないことを保証
- 匿名化されていないデータだからこそ気づく残差の系統変動

	都道府県	市区町村	残差
• 1113	大阪府	大阪市	-17.5
• 643	東京都	豊島区	-17.3
• 631	東京都	新宿区	-16.1
• 641	東京都	中野区	-15.8
• 637	東京都	目黒区	-13.9
• 642	東京都	杉並区	-12.7
• 638	東京都	大田区	-12.5
• 636	東京都	品川区	-12.2

婚姻件数と出生件数との散布図
出生件数 $\propto$ 婚姻件数！



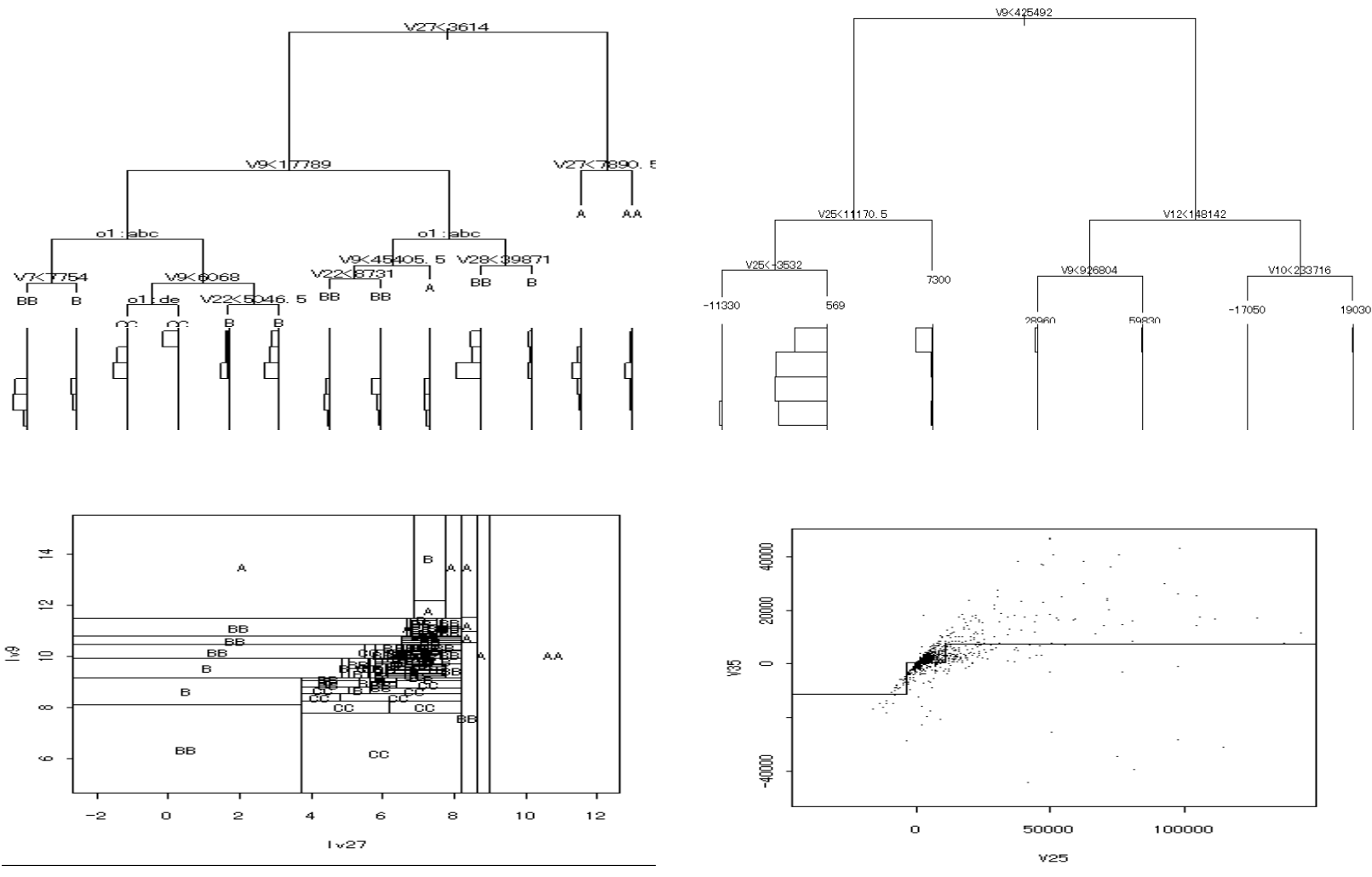
仮説成長型データ分析の手順と基礎概念

筑波大学経営システム科学専攻多変量解析

1997-2007で教えていたこと（椿,1999）

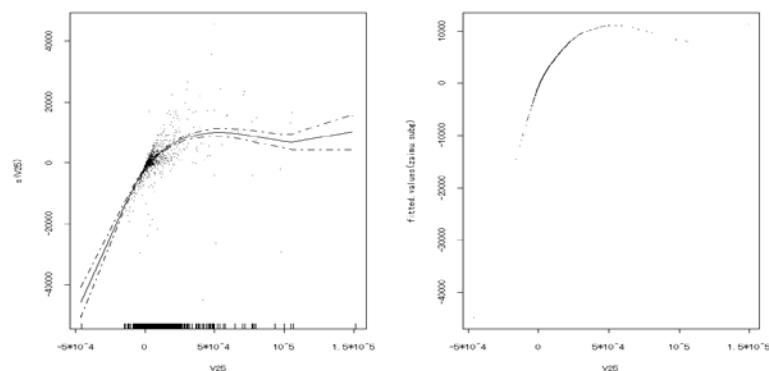
- 起)曖昧な仮説の提示（原始仮説）
- 承)樹形モデル（層別原理）当てはめでif-then規則型仮説への成長
 - 交互作用、例外構造の探索
- 転)一般加法モデル(平滑化原理)当てはめでパターン型定量仮説への成長
 - 非線形構造の探索
- 結)数式モデル(一般線形モデル等)の当てはめで、メカニズム型仮説への成長

承：If-then ruleからの出発



転：要因効果のパターン抽出

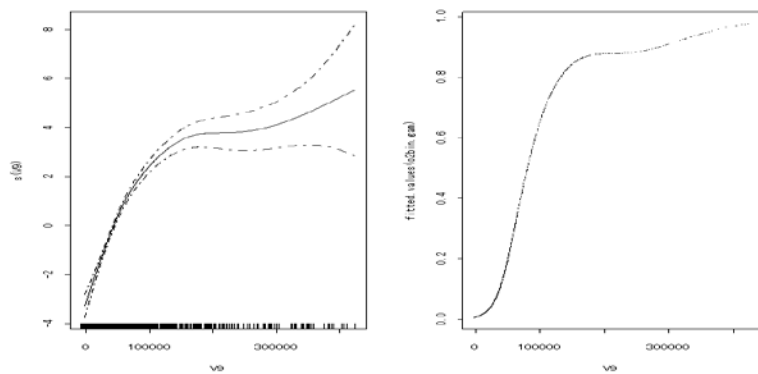
散布図平滑化



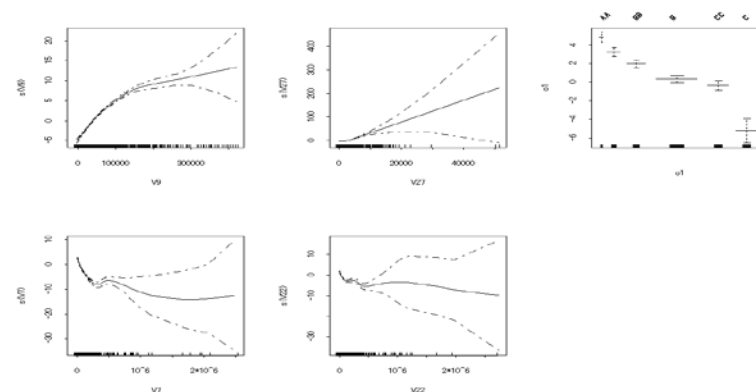
散布図を
滑らかに
繋ぐのが
本来の回帰関数

回帰関数：
条件付期待値関数

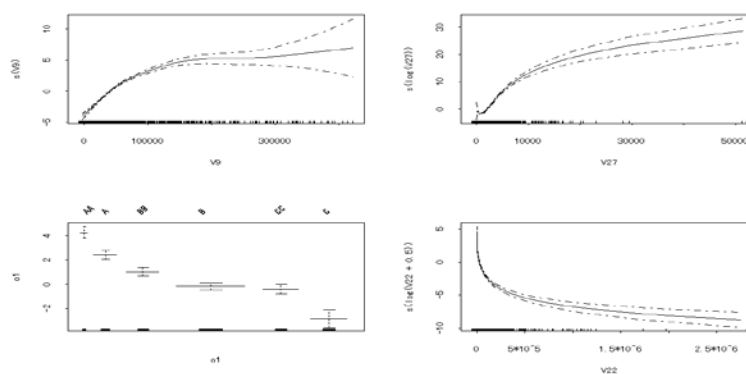
質的予測の場合



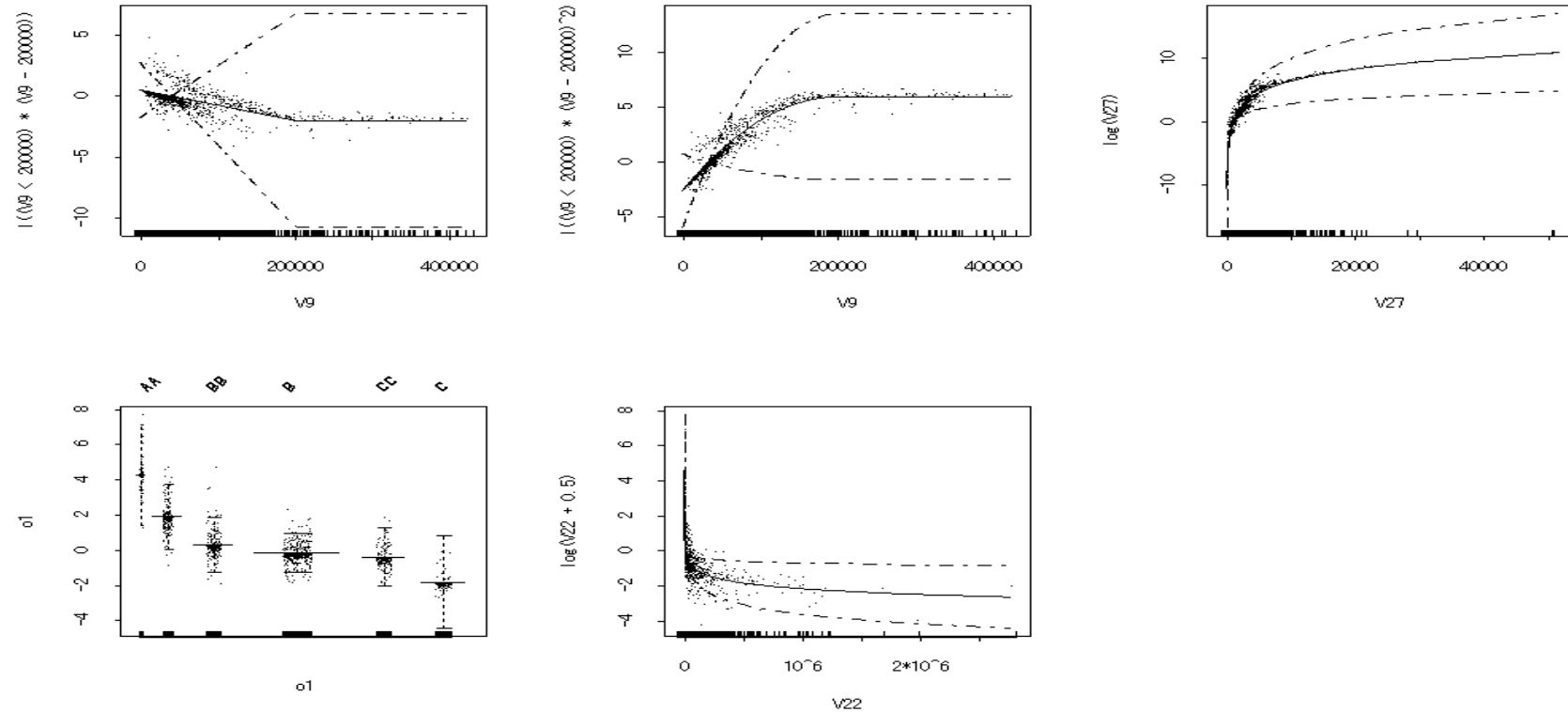
安全性への要因効果



変数変換による安定化



結：要因効果の数式化



- ここには、本来「固有科学」の知が必要：数理なき実証！
- **現在は少し違う考えも**，少なくとも初中等統計教育では持ちつつある

期待：経験からの帰納と理論からの演繹のバランスを：数理のある実証

田口の理想機能アイデアへの先祖帰り初等統計教育としての真っ当さ？

Gilchrist, W. (1984) *Statistical Modelling*, Wiley

- 学校と自宅との直線距離 x と通学ルート y の全長を調べ、その関係をモデル化せよ
- **Gilchristの実験例：Sheffieldのランダムな20地点対の場合**

Road distance	Linear Distance	Road distance	Linear Distance
10.7	9.5	11.7	9.8
6.5	5.0	25.6	19.0
29.4	23.0	16.3	14.6
17.2	15.2	9.5	8.3
18.4	11.4	28.8	21.6
19.7	11.8	31.2	26.5
16.6	12.1	6.5	4.8
29.0	22.0	25.7	21.7
40.5	28.2	26.5	18.0
14.2	12.1	33.1	28.0

理論に基づく演繹的思考＋法則の誤差 次のような演繹的議論を順次行う

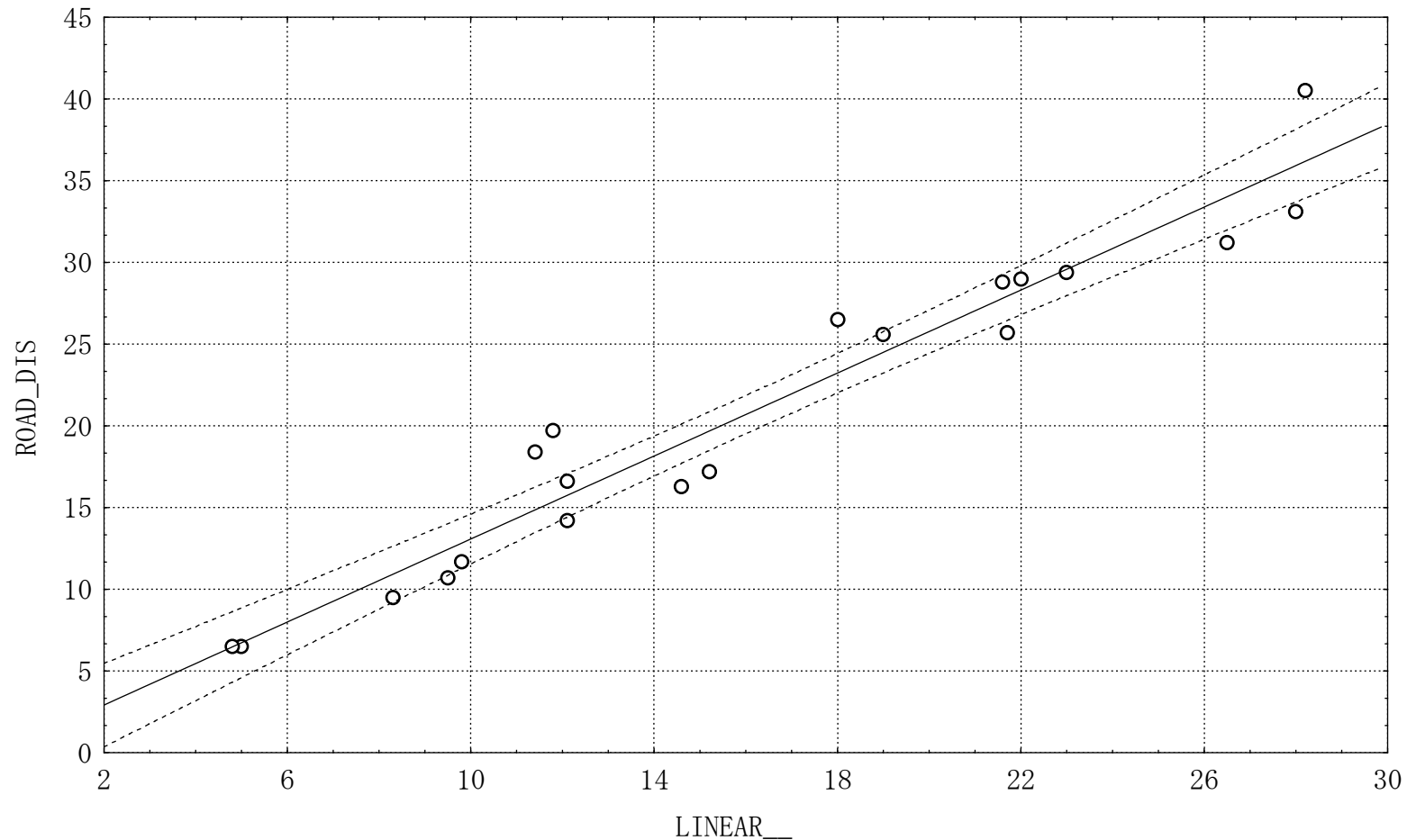
- 成立すべき厳密な数学的性質と不確実性
 - A) $x=0$ ならば $y=0$
 - B) $y \geq x$
 - C) x が増大すれば y も増大する傾向
 - 予測不可能な不確実性、すなわち偶然成分 e が存在
 - これは、道路（通学路）パターンのランダム性に起因
 - D) 上の考察を進めて、 x が倍になれば、近似的に y も倍近くなるという仮説も有力
- モデルの候補と誤差を含む演繹的考察の整合性
 - a) $y=x$
 - b) $y=x+e$
 - c) $y=x + \text{定数} + e = x + \mu + e$
 - d) $y=\text{定数} \times x + e = \theta x + e$
 - どのモデルが、演繹的思考と整合的となるか
 - A,B,C,D ⊢ a、b、c、dとの整合性をチェックしなさい

帰納的（経験的）考察

1 次式を当てはめるモデル ニ）が支持
直観で線を引けば $y = 1.3x$ といったところ

散布図 (S295. STA 10v*20c)

$$y = 0.379 + 1.269 * x + \text{eps}$$



初等的な統計的機械学習 初中等教育の可能性

初中等統計・数理教育と人工知能

CARTの層別原理

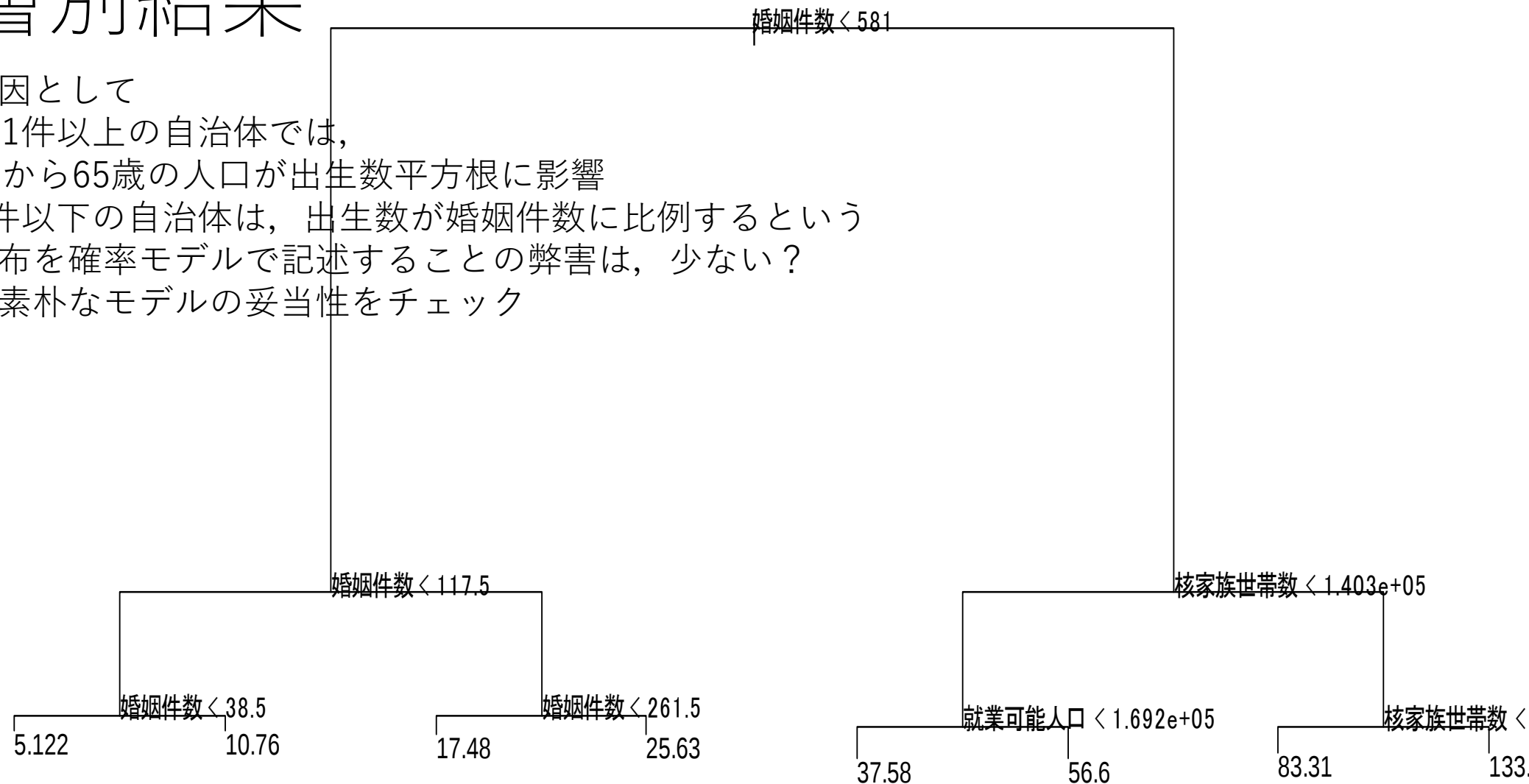
層別(stratification)と その最適化：人工知能

- 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くと有利である。
 - 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：その単純な最適化問題
 - 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→エントロピーと同等

CARTによる出生件数平方根の 自動層別結果

出生数に関する要因として

- 婚姻件数が年間581件以上の自治体では、
核家族世帯数や15歳から65歳の人口が出生数平方根に影響
- 婚姻件数年間581件以下の自治体は、出生数が婚姻件数に比例するという
統計モデルの残差分布を確率モデルで記述することの弊害は、少ない？
⇒人工知能系手法で素朴なモデルの妥当性をチェック

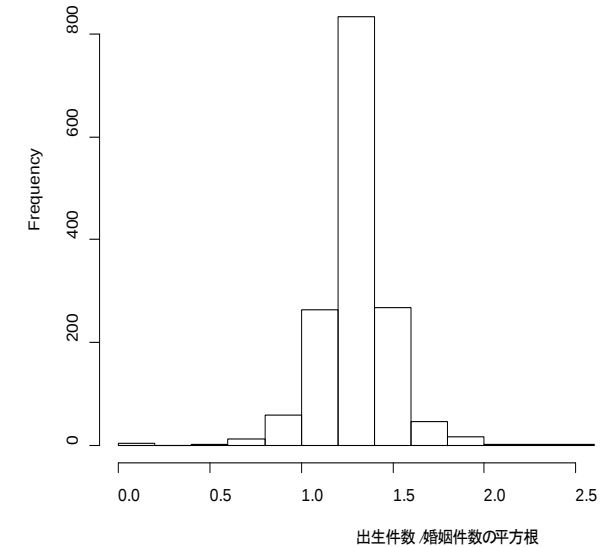
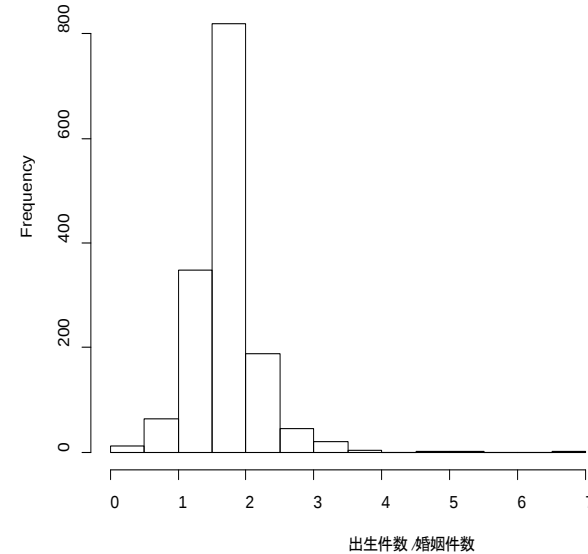


効率最適化に関する推論

効率が定義できるということ

効率指標の導出可能性と要因解析

- 出生件数と婚姻件数とが近似的に比例関係
- $\Rightarrow$ 出生件数／婚姻件数：効率指標
 - 正規分布に近づく変換は？
- **効率最適化**に対する統計的考察
 - 効率の外れ値分析
 - 出生件数/婚姻件数が4以上となる自治体
 - 北海道愛別町，山口県阿武町，
 - 宮崎県木城町
 - 効率に影響を与える要因についても効率データに変換



出生件数と婚姻件数の比のヒストグラム
右図は，平方根変換で正規分布近似

効率の最大化に資する要因の 人工知能による探索



- 説明変数に比率指標を追加
 - 女性65歳以上人口比率が大⇒
 - 婚姻当たり出生件数平方根が増大

女性高齢者比率の婚姻当たり出生平方根への影響

おわりに

- 初中等統計的方法教育の目指すところについて私見
 - 人工知能（統計的機械学習）的アプローチ：
⇒その背景数理（二次式の最適化）は初中等教育でも講義可能
 - しかし，初中等統計教育の本道か？
 - 統計モデルは，数理的裏付けのある比例関係など数理モデリングを修飾する方法と位置づけ？
 - 田口の理想機能（信号因子との比例関係）
- **統計的機械学習と古典的統計との関係性については，
水曜午後のNair教授の特別講義（積水化学京都研究所[科搜研]講堂）に期待**
- 一定規模の生徒・教員が誰もが理解できるデータ
 - 生徒の統計的推測，モデリングに関する多様な活動の可能性
 - 今後，種々の事例が蓄積し共有されることを期待
- **統計的方法教育の背後にある本質：科学的方法論（メタ方法）の教育**
 - **ICOTS10で扱う主要話題**