オンサイト拠点の活用について
ー提供者視点から利用者視点へー

統計数理研究所
椿　広計

内容

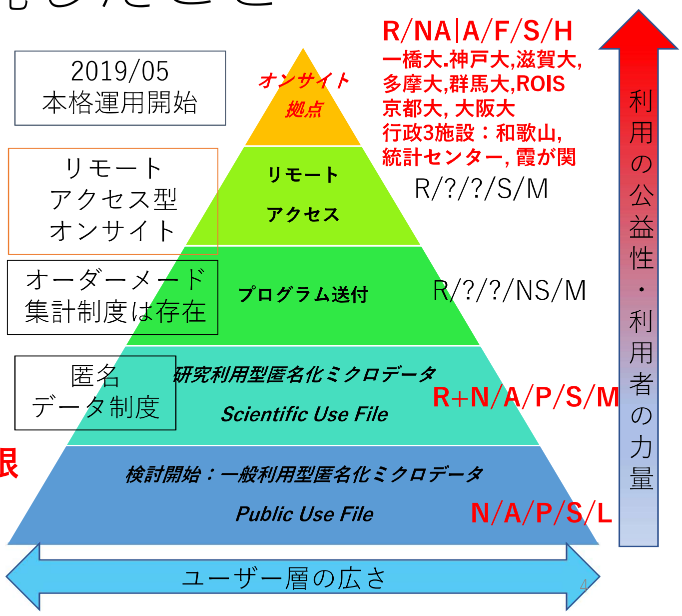
- はじめに：制度提案の立場
  - オンサイト拠点の本格駆動まで
- 統計センター時代の試行利用経験
  - 和歌山データ利活用センター試行事業：匿名データ
  - 自殺総合対策：国民生活基礎調査の目的外申請
  - 統計研究研修所の分析室：労働力統計
- おわりに：統計数理研究所に戻ってからの利用者経験
  - 現在進行形のプロジェクトでオンサイト拠点使ってみて感じること

はじめに：制度提案の立場

オンサイト拠点本格駆動（2019年5月）まで

公的統計ミクロデータ利活用
日本が実現したこと

- *提供データの分類軸*
  - R:**実データ** vs N:ノイズ付加データ
  - A:匿名データ vs NA:非匿名データ
  - F:**完全データ** vs P:一部データ
- *データを直接眺められるか？*
  - **S**, NS
- *提供サービスのAccessibility*
  - L:常時・誰でも・何処でも利用可能
  - M:利用者・利用期間の制限
  - H:利用者・利用期間・**利用場所の制限**
    - 分析結果の持ち出し審査



これまでの道のり：
人間・社会研究における公的統計活用

- 1990年代の状況
  - 国際競争力を低下させる日本の社会科学研究
  - 1980年代欧米で大きな変化
    - 自国公的統計ミクロ（個票）データの研究利用が可能に
    - 経済・社会研究の中心が集計データからミクロデータ分析にシフト
  - 日本の研究者の危機感
- *新統計法下のミクロデータ研究利用の先駆け研究*
  - 科研費特定領域:1996-1999
  - 統計情報活用のフロンティアの拡大の総括的研究:松田芳郎（一橋大）

- 13研究班の要請を集約：5省庁17調査について、総務庁長官に「目的外使用申請」
- データ処理センターとして活動
  - オーダーメード集計制度の先駆け
- リサンプリングデータの作成
  - 匿名データ提供制度の先駆け
- イミテーションデータの作成
  - 一般ミクロデータ制度の先駆け
- データ処理結果および処理要求をデータの秘密保持を保ったうえで各研究班相互を
計算機ネットワーク化して計算結果等の情報を流通させるシステムを開発
  - 現行のリモートアクセス型オンサイト拠点制度の先駆け
- 日本評論社からの「ミクロデータ分析シリーズ」により啓発

統計データの二次利用促進に関する経緯

- 2007年総務省政策統括官室
- 統計データの二次利用促進に関する研究会
  - 廣松毅座長：内閣府統計委員会へのInput
  - 統計法改正に向けた二次利用のスタイルと展開
    - *オーダーメード集計・匿名データ・疑似ミクロデータ*の提供
    - 利用目的としての公益性
- 2009年統計法公布後の活動
  - ミクロデータの提供方法
    - 各国制度比較と日本の特異性
    - 従来の目的外申請
      - 個人情報・法人情報が付随したデータを
セキュアな環境を持たない研究者が管理する可能性

  - オンサイト拠点：各大学などが整備
    - 個票をセキュア監視環境下で分析
    - 探索的モデリングが可能
    - 各拠点ごとの人員・設備整備にかなりなコスト
  - リモートアクセスによるオンサイト拠点での分析ネットワーク形成
    - 各拠点にはデータは置かない：中央で一括管理、事前審査から持出し審査へ
- 2013年：川崎茂応用統計学会長
  - 日本学術会議マスタープラン提案
  - 採択
➡内閣府統計委員会
基本計画部会次期基本計画への反映

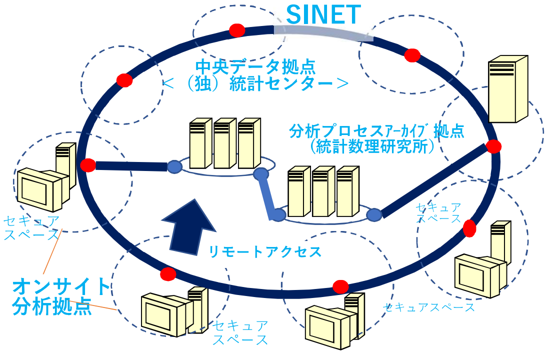
学術会議マスタープラン：2013

目標

- 国の保有する公的統計のミクロデータを実証研究に継続的に活用することのできる分析ネットワークを全国規模で構築することにより，
  我が国の人文社会科学における実証研究を質・量の両面において飛躍的に発展
- 社会・経済に関する実証分析の発展を通じて，**我が国の公共政策における「事実に基づく政策形成」（Evidence-Based Policy Making)の普及に貢献**

具体的内容

- ミクロデータ提供の中核となる**「中央データ拠点」**を**総務省/（独）統計センター**に整備。
- ミクロデータの分析拠点として，各都道府県につき最低一つの大学等にセキュリティを確保した**「オンサイト分析拠点」**を設置
- 研究成果の事後検証・再利用に資するため，統計数理研究所に**「分析プロセスアーカイブ拠点」**を設置。
- セキュリティ確保のため，**SINET**により拠点間を接続

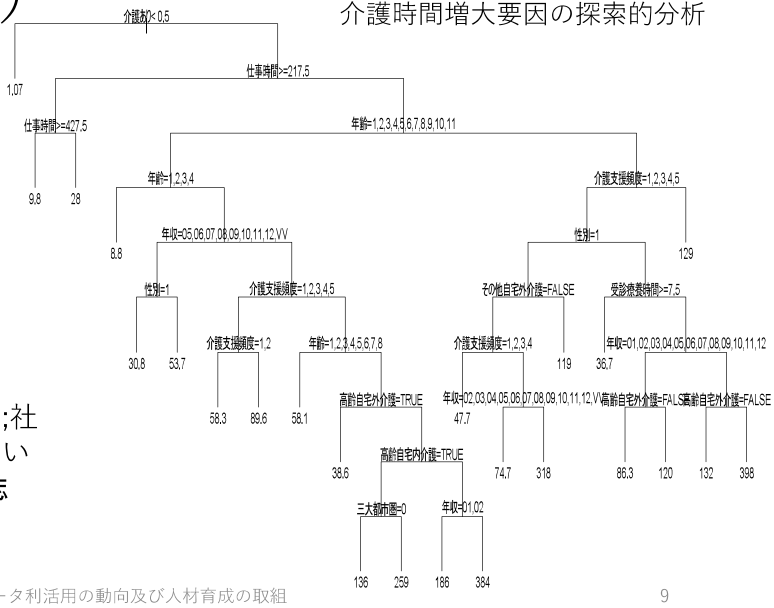


統計センター時代の試行利用経験

和歌山データ利活用センター試行事業：匿名データ
統計研究研修所の分析室：労働力統計

ミクロデータ探索分析に期待すること（匿名データ利用）

介護時間増大要因の探索的分析



- 和歌山実証実験(2016/07)
  - ミクロデータ官学活用拠点
    - 「統計データ利活用センター」
    - 和歌山県設置可能性
  - 1週間和歌山市で
    - 社会生活基本調査匿名データ分析
      - 大井達郎（和歌山大学）
      - 岡檀（和歌山県立医大、当時）
    - 統計センターとして分析支援
  - 探索的データ解析のEBPM
- 岡、山内、椿(2017)
  - 在宅介護負担が増える要因の探索;社会生活基本調査・生活時間編を用いての検討,**日本社会精神医学会雑誌** 26(3) 246 - 247

椿、会田(2019) リサンプリングによる労働力調査推定精度評価、統計研究研修所彙報 Vol.76, pp. 39-50.

- 我が国の就業・不就業の状況を把握するため，
  一定の統計上の抽出方法に基づき選定された全国約4万世帯の方々を対象に毎月調査
  - 1947年7月から本格開始：全国1万5000世帯：層化3段抽出
    - 米国労働力調査参考：標本理論の適用
  - 1950年：指定統計30号
  - 1954年：現行の層化2段抽出
  - 1961年：標本サイズ25000世帯
    男女・年齢階級・地域別の毎月末日15歳以上推計人口を
    ベンチマークとし，線形推定を改善する比推定採用
      - 比推定用乗率 ＝ベンチマーク人口 / 人口の線形推定値
        - 労働力調査ミクロデータに集計用乗率として付与
  - 1983年：全国4万世帯約10万人を対象：地域別四半期推定

調査区リサンプリング方法
2通りの比較実験

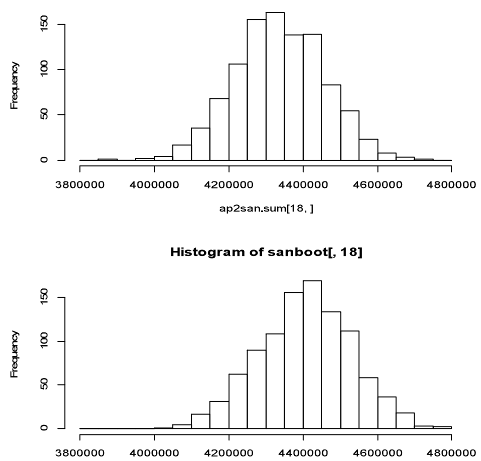
- 推定精度評価は，２段階のブートストラッピング
  - 第１段階
    - **抽出調査区のリサンプリング**
  - 第２段階
    - リサンプリング調査区内の世帯のリサンプリング
- ミクロデータにない補助情報が必要な方法
  - 方法１：抽出調査区から抽出確率に比例する確率で復元抽出
- ミクロデータで可能なリサンプリング方法
  - 方法2：抽出調査区から単純無作為復元抽出
- 推定量の調査区・世帯2段階復元抽出毎に集計結果を算出し、その分布（ブートストラップ分布）を比較

リサンプリング実験

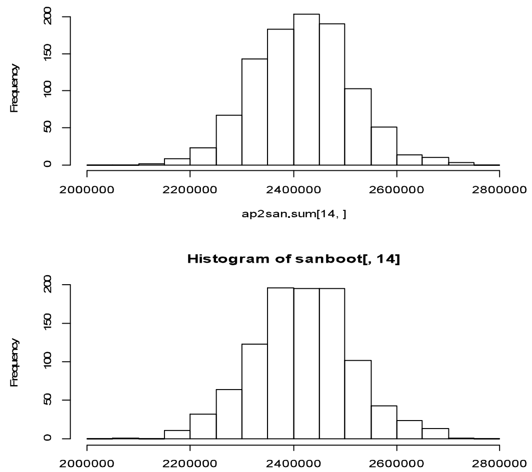
- 利用データ：労働力調査2017年9月集計対象
  - ミクロデータ利用申請
    - 総務省統計研究研修所内でリサンプリング実験：曾田，椿
      - データはフルセットでなく、必要最低限度
        - **メモリーは椿科研で8GBに拡張：全く動かない**
    - **別途調査区の抽出に関わる情報を総務省統計局より提供**
- 単純な線形推定量のブートストラップ分布の導出：
  - リサンプリング回数=1000 :調査区自体も再抽出
- 利用プログラミング言語：R
- 対象：産業別就業者数

リサンプリング分布の比較
上段：忠実，下段：単純

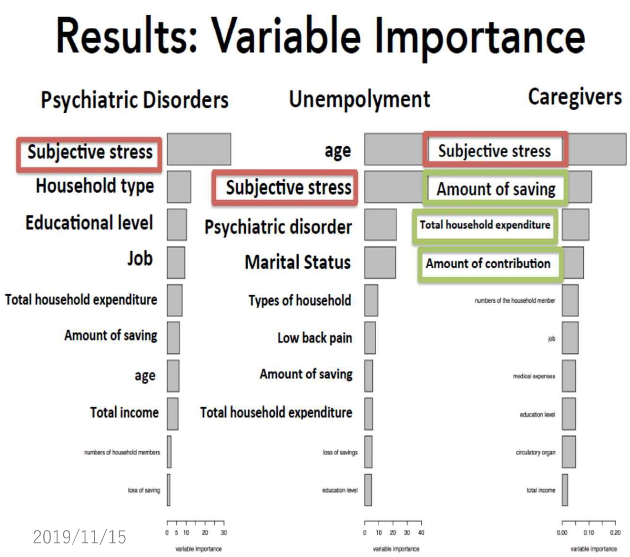
他に分類されないサービス産業の就業者数推定値

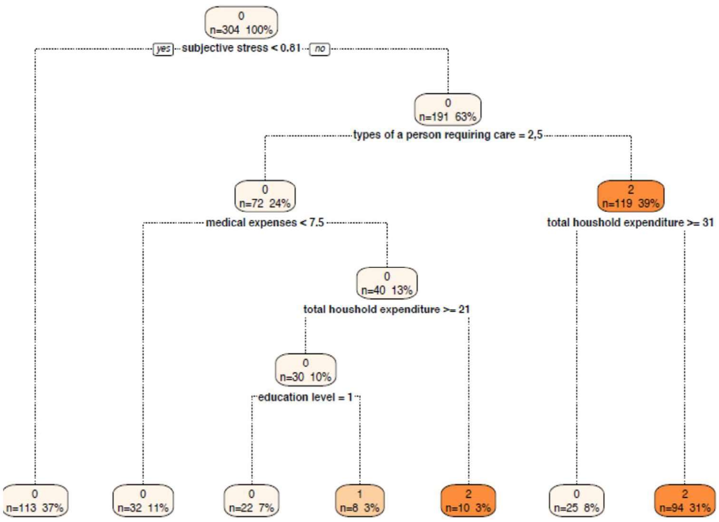


生活関連サービス産業就業者推定値



Takebayasi, Kubota and Tsubaki (2016)
Risk profiles for severe mental health problem
The 22th international conference on computational statistics
国民生活基礎調査匿名データ





介護者のK6悪化要因の分析

厚生労働科研自殺総合対策研究の場合 平成30年度分担研究報告書より

- 本研究は、自殺総合対策に資する公的統計データの利用環境を各府省と連携して構築整備すること、さらに具体的にその種のデータを利活用して、自殺対策に資する実証研究を加速することを目的としている。
- 本研究は、平成27年度に実施した国民生活基礎調査（注：匿名データ利用）K6に対するリスク分析を基に、平成28年度厚生労働省に対して国民生活基礎調査ミクロデータのK6情報を地域政策に資するために行うことを研究目的とした。
- しかし、地域情報とのリンクも可能にするという統計法33条に基づく目的外申請が平成28年度不調に終わったため、別途整備を進めたオンサイト拠点において、自由度の高い探索的な分析を実施する事、さらにはオンサイト拠点において厚生労働省ミクロデータを利活用できるようにすべく、研究目的を変更した。
    - 中略
- 公的統計ミクロデータの利活用については、最も自殺総合対策に資すると総務省統計審議会（旧審議会）でも指摘されていた「国民生活基礎調査」のミクロデータ、特にK6に関する情報が、依然としてオンサイト拠点で利用可能となっていない。人口動態統計という極めて重要なミクロデータが分析可能になったことを契機に、統計データ理科利用センターを通じた厚生労働省との交渉を強化したい。

おわりに： 統計数理研究所に戻ってからの利用経験

現在進行形のプロジェクトでオンサイト拠点使ってみて感じること

現在進行中のプロジェクト
今年の8月から、私も実際に統数研オンサイト使い始めました

- 厚生労働科研　本橋　豊（自殺総合対策センター長）班
  - 分担研究者　椿 (ISM)
  - 研究協力者：久保田(多摩大)、岡（ISM）、
    岡本（ISM）、竹林（福島県立医大）
  - **分担研究テーマ**
    - **エビデンスに基づく自殺問題の総合対策の確立に向けて**
      - **社会生活基本調査からリスク要因を抽出**
        - **和歌山試行実験データに地域情報を付与：地域特性の分析**
        - 市区町村マクロ人口統計情報、経済統計情報などを結合（完了）
        - 市区町村自殺率情報を結合（未完）
    - **介護時間増大リスク：本日岡檀先生から中間報告**
    - 自殺増大リスク要因（これから）

公的統計ミクロデータコンソシアム2017
統計センターからの期待に現在追加したいこと

- ミクロデータユーザー会としてのGood Practice共有
  - **もっとこの制度を知らしめなければならない**
- 政府・地方政府のEBPMを支援して欲しい
- オンサイト拠点の顧客として様々な無茶な要求をして欲しい
  - **申請、持ち出し審査は迅速に対処いただき大いに感謝**
  - **CSVファイルとして提供いただいていることも朗報（符号表からの解放）**
  - すぐに使えるミクロデータが欲しい
  - 情報のリンケージをどんどんやって欲しい **（どんどんやりたい）**
    - **市区町村、県のマクロデータはオンサイト環境に常時置いておけないか？**
  - オンサイト拠点ではなく自分の研究室で分析したい
  - まだ公開されていないあのデータを分析したい
    - **国民生活基礎調査が使いたい**
  - パブリックユーズファイルはこういうものであってほしい
  - **仮想PCのメモリー2GBでは非力**
    - **社会生活基本調査：エクセルで読むとフリーズ**
    - **分析途中でワードの立ち上がなどでフリーズ**