経済統計・政府統計の理論と応用 2016 [1]

山本拓 [2] & 国友直人 [3] & 川崎能典 [4]
(共編)

2016 年 2 月

[1]文部科学省・科学研究費プロジェクト「経済統計・政府統計の理論と応用からの提言」(2015 年度〜) が 2016 年 1 月 29 日に東京大学経済学研究科において開催した研究集会における講演をまとめたものである。
[2]統計研究会
[3]東京大学大学院経済学研究科
[4]統計数理研究所

前書き

本報告書は、日本学術振興会・科学研究プロジェクト「経済統計・政府統計の理論と応用からの提言」（2015 年度－ 2018 年度、研究代表者：山本　拓）が、2015 年 1 月 29 日 (金) に東京大学小島ホールにおいて開催した 2015 年度の研究集会における講演内容をまとめたものである。なお本研究集会のオーガナイザーは、川崎能典、国友直人、山本拓が務めた。

本プロジェクトの目的は、経済統計・政府統計における主要な課題の、技術的および制度的問題を、統計学的な立場から理論的・学術的に検討し、具体的解決策を提言することである。経済統計、とりわけ政府統計は、経済・社会の動向を理解し、政策を実施、評価するためには不可欠な情報であることは言うまでもない。最近は evidence-based policy ということもよく言われ、政府統計の重要性は一般に広く認識されつつあると思われる。しかし、経済統計・政府統計への信頼性は、近年必ずしも増しているとは言えない状況である。経済社会の急激な変化に伴い、政府統計の質の確保が困難になりつつある。マクロ経済統計の側面では、GDP 統計などに代表されるマクロ公表系列の質と信頼性の問題、信頼性の高い将来人口の推計の問題、地域による経済情勢のばらつきの把握などの問題を挙げることができる。またミクロ経済データにおいては、統計調査をとりまくプライバシー意識の高まりから、調査精度の確保が難しくなりつつあるという問題や、情報開示と秘密保持の両立という匿名化問題などを挙げることができる。新しい統計学的知見の導入に関しては、日本の政府統計部局が分散化されているために、これまでは、個別の担当部局あるいはその時々の担当者に個別に招かれた研究者によって知見や助言が提供されることが多かった。政府統計を巡る重要な論点について、担当部局をまたいでその知見が共有されることは少なかったと思われる。またそれらの話題が広く研究者間で議論されることも少なかった。そのような意味で、経済統計・政府統計の技術的・制度的問題点を、統計学的立場から総括的に検討していくという本研究プロジェクトは、一つの新しい方向性を目指したものである。

本プロジェクトの研究集会は、プロジェクトのメンバーと実際に経済統計・政府統計に作成者または利用者として携わっている方々との積極的な交流を提供することをその重要な目的の一つとしている。今回は本プロジェクトの第 1 回目の研究集会ということで、メンバーが今後の研究上の刺激を得ることを期待し、外部の方の報告が多くなるように構成した。

第 1 セッションでは、政府統計・経済統計を巡る最近の重要な問題が扱われた。すなわち、政府統計の役割として重要な位置をしめる 2 次利用の現状と課題、長期的な地方創生に関して、重要な要因となる府県別の人口変動の予測、さらにマクロデータとミクロデータの融合の方法についての新しい試み、が報告された。第 2 セッションでは、政府統計・経済統計に関わるかねてからの問題に対する理論的と応用からの接近である。すなわち、匿名データの開示リスクの評価の考え方と応用例、ならびに新しい季節調整法 X13 の特徴の解明とその評価である。第 3 セッションは、近年のその重要性が際だってきた小地域統計の問題に的を絞ったものである。さまざまな具体的な応用をふまえた理論的問題が明らかにされ、それらについての新しい理論的接近法が提案され、その応用可能性が報告された。

このような研究集会が情報交換ならびに刺激となり、経済統計・政府統計の今後の改善の一助になることを期待する次第である。

2016 年 2 月
編者

研究集会・プログラム

科研プロジェクト「経済統計・政府統計の理論と応用」
日程：2016 年 1 月 29 日 (金)
会場：東京大学経済学部小島ホール 2 階
オーガナイザー：山本拓・国友直人・川崎能典

＜挨拶＞
13:00〜13:05「研究プロジェクトの計画」山本拓

＜セッション I＞ 政府統計・経済統計を巡る課題
Chair: 川崎能典

13:05〜13:40「公的統計 2 次利用などに関わる取り組み」椿広計 (統計センター)
13:40〜14:20「地方創生と人口統計」金子隆一 (社会保障人口問題研究所)
14:20〜15:00「マクロデータとミクロデータの統計的データ融合について」星野崇宏 (慶応大学)

＜休憩＞

＜セッション II＞ 政府統計・経済統計の理論と応用
Chair:国友直人

15:10〜15:50「匿名データの開示リスク評価例」星野伸明 (金沢大学)
15:50〜16:20「季節調整プログラム X-13ARIMA-SEATS について」高岡慎 (琉球大学)

＜セッション III＞ 小地域統計の理論と応用
Chair:久保川達也

16:20〜16:50「小地域推定問題に対する”モデルに基づくアプローチ”の新たな課題—海外の事例を通して–」廣瀬雅代 (統計数理研究所)
16:50〜17:20「正値地域データを解析するための変換モデルについて」菅澤翔之助 (統計数理研究所)
17:20〜17:50「空間重み付き経験ベイズ推定と死亡データへの応用」川久保友超 (千葉大学)

公的統計2次利用などに関わる取り組み

独立行政法人　統計センター
椿　広計

統計センターの紹介：３つのミッション

国民の合理的な意思決定に必要不可欠な公的統計を正確に作成し、
遅滞なくかつ効率的に提 供するため、統計センターに対して次の使命を果たすことが求められている。

１．統計をつくる

人口や失業率、消費者物価指数等の我が国の基幹的な統計の作成

２．統計を活かす

統計利用者、調査対象者、研究者が便利に安心して活用できる統計サービスの提供

３．統計を支える

各府省、地方公共団体、国際機関、各国政府等の統計作成を支えるシステムの運用管理やプロジェクトの遂行

所属組織
独立行政法人
統計センターの
最近の活動
「統計を活かす」

JSTAT MAPは、
事業所に
多様な近傍を
定義し、性別年齢
別人口分布を
すぐに表示

ある研修:
浦安市の介護
施設配置の
あるべき政策？

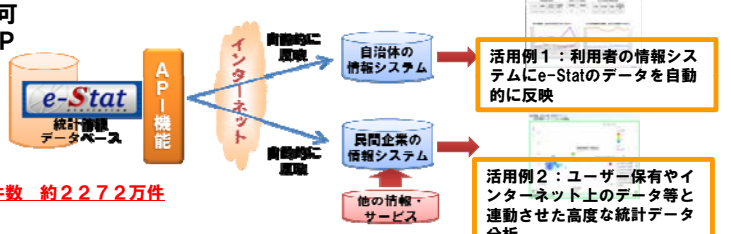
統計におけるオープンデータの高度化

統計データの提供方法を高度化し、新たな付加価値を創造するサービスや革新的な事業の創出などを支援する取り組みを、総務省統計局と連携し行っており、政府が取り組んでいるオープンデータの推進を先導。

ＡＰＩ機能による統計データの提供

2014.10.31から運用開始

統計データを機械判読可能な形式で提供するＡＰＩ機能（Application Programming Interface）を提供中



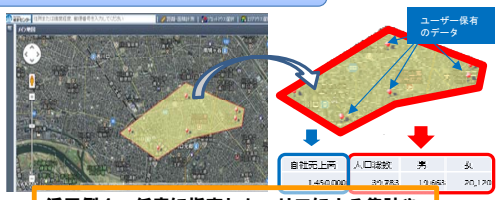
開発支援情報も提供中

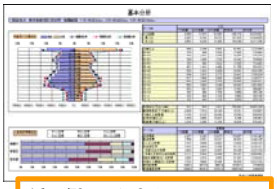
利用登録者数　３１６２人
統計データへのリクエスト件数　約２２７２万件
（平成27年3月26日現在）

地図による小地域分析（jSTAT MAP）

2015.1.20から運用開始

任意に指定したエリアによる集計や利用者が保有するデータの取り込み集計する機能などを提供





タブレット用アプリも提供中

利用登録者数　２３３４人
ログイン件数　約２３０００件
（平成27年3月26日現在）

活用例１：任意に指定したエリアによる集計や、利用者が保有するデータと統計データを組み合わせ、集計結果を地図上で視覚的に把握可能

活用例２：選択したエリアの年齢構成等の基本的な分析結果のレポート作成

Needs: 人間・社会研究における公的統計活用

2009年統計法全面改訂

- 旧統計法（1947年）法の目的
  - 第一条 この法律は、**統計の真実性**を確保し、統計調査の重複を除き、統計の体系を整備 し、及び統計制度の改善発達を図ることを目的とする。
- 新統計法第一条
  - この法律は、**公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な 情報**であることにかんがみ、公的統計の作成及び提供に関し基本となる事項を定めることにより、公的統計の体系的かつ効率的な整備及びその有用性の確保を図り、もって**国民経済の健全な発展及び国民生活の向上に寄与**することを目的とする。
    - オーダーメード集計や匿名データの作成・研究目的などでの提供を可能とする統計データの二次的利用の制度設立

統計データの二次利用促進に関する研究会

- 総務省政策統括官室：2007年スタート
  - 廣松毅（東京大学，現情報セキュリティ大学院大学）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
- 統計法公布後の活動
  - 提供方法：各国制度比較と日本の特異性
    - 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
    - 目的外申請で得た個票をセキュアな監視環境下で分析
    - 探索的なモデリングを可能とする。しかし，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ネットワーク形成：応用統計学会からの学術会議マスタープラン提案
    - 各拠点にはデータは置かない：監視は中央で一括
  - 統計データ・アーカイブの整備

匿名データの提供

一般からの申出を受け、利用要件を満たした申出者に対し、特定の個人又は団体等が識別できないように加工して作成した調査票情報の利用を一定期間認める制度。匿名データを利用することで、行政機関が作成していない統計表の作成のみならず、多変量解析などミクロデータに基づく実証分析を行うことが可能。

【利用要件】
- 統計の作成または統計的研究にのみ利用されること
- 学術研究目的または高等教育目的の用に供することを直接の目的とすること
- 学術研究の成果または高等教育の内容が公表され、社会に還元されること
- 匿名データが適切に管理されること

オーダーメード集計

一般からの委託を受けて、利用要件を満たした申出者に対し、調査票情報を用いて集計を行い、その結果の提供を行う制度。オーダーメード集計を利用することで、行政機関等が作成していない統計表に基づいた分析が可能。

【利用要件】
- 統計の作成または統計的研究にのみ利用されること
- 学術研究目的または高等教育目的の用に供することを直接の目的とすること
- 学術研究の成果または高等教育の内容が公表され、社会に還元されること
- **⇒平成28年度から民間においても研究に資するものへ拡大**

擬似ミクロデータの試行的提供

擬似ミクロデータは、我が国の行政機関が実施した統計調査の集計表から作成したミクロデータ形式の擬似的なデータで、大学等の教育機関における授業や演習及び公的統計の二次的利用の際のテストデータなどの利用が可能であり、ミクロデータ利用者の裾野を広げ、公的統計の二次的利用の拡大を図るため無償で提供している。

【利用要件】

- 申請者及び利用者以外の者に利用させないこと
- 集計表から擬似的に作成したデータであるので、分析結果は実証研究の結果と見なすことができないことを理解すること
- 利用者アンケートを提出すること

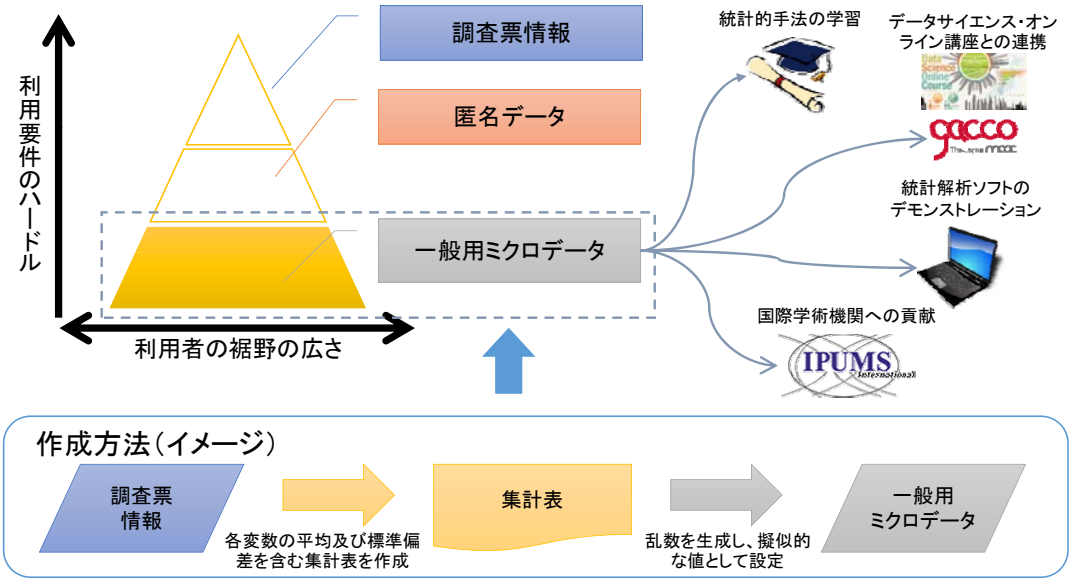
試行提供している擬似ミクロデータ（平成27年5月末現在）

- 平成16年全国消費実態調査に関する試行版の擬似ミクロデータ
  - 大規模データ （CSV形式のみ）
    - 約200項目：世帯情報、収入、実支出（消費支出、非消費支出）、実支出以外、繰越金
    - 約3万2千レコード
  - 簡易データ （CSV形式　及び　Excel形式）
    - 25項目：世帯情報、消費支出
    - 約8千レコード

3

今後の予定

擬似ミクロデータの提供は、平成27年度末で終了する。
平成28年度からは、一般用ミクロデータの提供に移行する。



統計センターと学術研究機関等との連携について

学術研究機関等との連携協力の取組

統計センターでは、平成21年４月施行の統計法において創設された公的統計の二次的利用制度の充実と学術研究の発展を図っていくため、学術研究機関等と連携した取組を展開。公的統計の利用拡大に係る取組に賛同する法人と連携協力協定を締結し、公的統計の二次的利用に関する研究・開発，普及・啓発を推進するほか、これらの法人に統計データアーカイブのサテライト機関の役割を担ってもらい、研究者等に向けた二次的利用のサービスの充実を図っている。

法人の要件

1. 国立大学法人法に基づき設置された国立大学法人及び大学共同利用機関法人
2. 私立学校法に基づく学校法人により設置された私立大学
3. 独立行政法人通則法及び個別法の定めるところにより設立された独立行政法人
4. 法人税法別表第1に掲げる公共法人
5. 公益社団法人及び公益財団法人の認定等に関する法律により公益性の認定を受けた公益法人（特例民法法人を含む。）

連携協力事項

〇公的統計に関するデータアーカイブの運営に関すること（施設基準に適合）
・匿名データの提供
・オンサイト利用環境の提供
など
〇公的統計の二次的利用に関する研究・開発
〇公的統計の二次的利用に関する普及・啓発
〇人材交流
〇その他協定の目的を達成するために必要な事項

連携協力協定を締結している大学等

サテライト機関名	匿名データの提供	オンサイト利用環境の提供	二次的利用に係るＵＲＬ
国立大学法人 一橋大学 経済研究所附属社会科学統計情報研究センター	○	○	http://rcisss.ier.hit-u.ac.jp/Japanese/micro/index.html
国立大学法人 神戸大学大学院 経済学研究科	○	—	http://www.econ.kobe-u.ac.jp/kuma/satellite/index.html
法政大学 日本統計研究所	○	—	http://www.hosei.ac.jp/toukei/micro/index.html
大学共同利用機関法人 情報・システム研究機構 新領域融合研究センター（統計数理研究所）	○	○	http://www.rois.ac.jp/tric/tokumei/tokumei.html

進行形の活動：公的統計マイクロデータ活用のためのネットワーク基盤

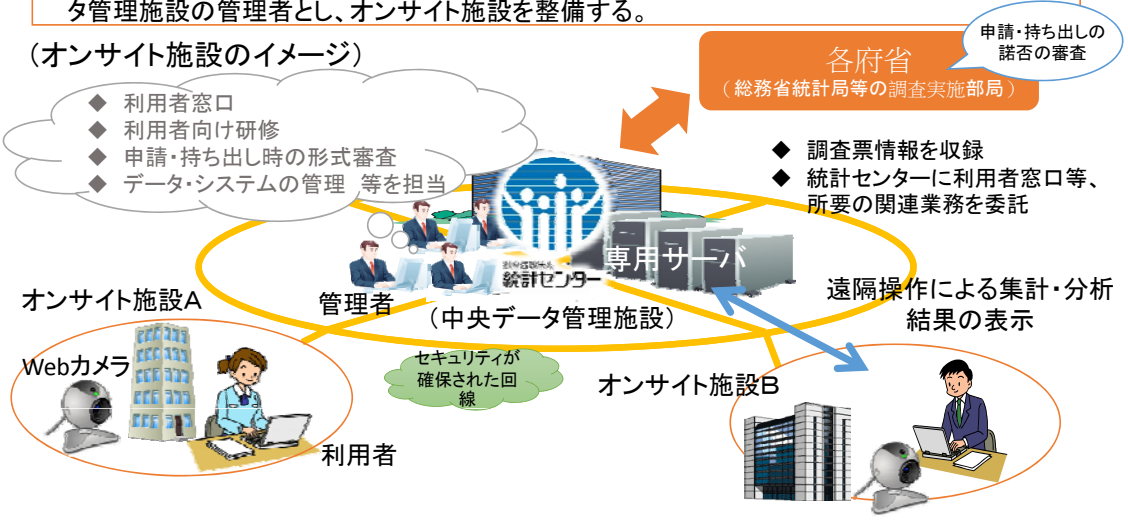
総務省「統計の2次利用研究会」
川崎茂応用統計学会長（当時）「学術会議マスタープラン」提出
内閣府統計委員会の「公的統計の整備に関する基本計画
基本計画2014年3月25日閣議決定

公的統計の整備に関する基本計画への組み込み

- 内閣府統計委員会
  - 2013年10月30日総務大臣より諮問，2014年1月31日答申
- 2014年3月25日閣議決定
- 調査票情報等の提供及び活用については、
  - セキュリティレ ベルや調査票情報等の匿名性の程度に応じた利用形態ごとの特性
  - 諸外国における取組状況等を総合的に勘案した上
- 法制度上の整理を含め、以下の取組を行う
  - オーダーメード集計における利用条件の緩和に向けた検討
  - 調査票情報の提供におけるリモートアクセスを含むオンサイト 利用やプログラム送付型集計・分析の実現に向けた整理・ 検討
  - 匿名データの作成及び提供における提供対象統計調査の種類や年次 の追加等によるサービスの充実
- その際、効率性及び利便性の観点から、政府一体として一元的な取組を推進する。
  - 主担当総務省政策統括官，府省横断の中核機関：(独)統計センター
  - 二次利用研究会での検討⇔リモートアクセス拠点形成⇔**大学などの公的統計ミクロデータ2次利用フォーラム形成**
    - **情報・システム研究機構平成28年3月29日一橋講堂で立ち上げ**：平成28年度概算要求：統計センター人員の確保

(参考)調査票情報のオンサイト利用(現在検討中)

○ オンサイト利用については、長年にわたる課題であったが、IT技術の進展等を踏まえ、平成28年度を目途に、リモートアクセスを活用したオンサイト利用の試行を開始する。今後も、総務省政策統括官(統計基準担当)は制度面の検討・関係府省と調整するとともに、総務省統計局と(独)統計センターが連携して技術面を検討する。
○ 具体的な試行については、統計局において、学界や各府省の協力を得つつ、統計局の統計調査の調査票情報を主たる対象に、(独)統計センターを政府共通の基盤である中央データ管理施設の管理者とし、オンサイト施設を整備する。



公的統計
ミクロデータ分析の課題

政府統計部局の役割
研究の障害除去：データがない、使いにくい

- 提供可能な公的統計ミクロデータの範囲拡大
  - 匿名データ・オーダーメード集計の範囲拡大
    - 事業所・企業統計などの匿名化困難
- 適切なメタデータの付与
  - 現状：統計局：CSVあるいはXMLでの提供
    - 研究者の負担
      - 研究に用いる複数データセットの変数結合＊国内外の類似データとの結合
  - 適切なメタデータの付与：**LOD**（Linked Open Data）形式によるデータ提供
    - LOD：関連するデータ同士が相互に結びついているデータ
    - **統計オープンデータモデル事業**
      - 福井県・統計局・統計センター協働事業
      - 統計局提供：国勢調査￥社会・人口統計体系等の統計データ
      - 福井県提供各種行政データ

調査環境の悪化と統計精度向上の社会要求

- 未回答
  - あってはならない未回答率の増大
  - Imputationの増大（回答データとして扱う問題）
- 欠測値だけのImputation（モデルベースの予測値）か？
  - 無作為抽出と不偏推定の意味
    - 抽出(観察)データの実測値Y_t
    - 非抽出データのモデルベースの予測値$Y(x_{t-1})$：Best Available Predicts
    - データの観測確率：$P(z_{t-1})$：抽出確率×未回答確率モデル
    - 不偏推定量の構成
      - 実測した場合：$Y(x_{t-1})+\{1/ P(z_{t-1})\}\{Y_t -Y(x_{t-1})\}$
      - 実測しない場合：$Y(x_{t-1})$

おわりに

- 社会のための科学の推進
  - 一学会、一大学、一研究所でできることの限界
  - 多様な価値を受容したネットワーク形成は必然
  - 産官学の連携
- 社会のための科学の支援基盤
  - データ基盤：公的統計のみならず重要
    - そのセキュリティ
  - 知識基盤：
    - モデル・プロセスの記述標準・産官学の共有・相互評価
- この種の事業を進める人財の価値評価

参考資料1

- 匿名データの利用に関するFAQ：
  - http://www.nstac.go.jp/services/faq-anonymity.html
- 擬似ミクロデータの利用に関するFAQ
  - http://www.nstac.go.jp/services/faq-gijimicro.html

（参考資料２）匿名データの利用実績例

利用目的	調　査　名	研究の名称
学術研究目的	社会生活基本調査	正規雇用者における平日の労働時間と休息時間 ―「社会生活基本調査」ミクロデータによる分析―
		生活行動からみる高齢者の行動特性について ―社会生活基本調査の匿名データを用いて―
		子供のいる世帯における夫と妻の2次活動時間の差異について ―社会生活基本調査の匿名データを用いて―
		趣味・娯楽活動の時間について
		個人・世帯属性と行楽・観光旅行行動の関係
	全国消費実態調査	等価尺度の推計と比較－消費上の尺度・制度的尺度・OECD尺度－ 『季刊社会保障研究』Vol.48 Spring 2013 No.4
		所得格差変動の年齢階級別要因分解 ：全国消費実態調査ミクロデータを用いて
	全国消費実態調査 住宅・土地統計調査	持家取得における既婚女性の就業の役割
	就業構造基本調査	転職経験および転職理由と転職希望意識との関連について ―就業構造基本調査匿名データによる統計分析
		若年者就業率における賃金弾力性の推定
		女性事務職の賃金と就業行動 －男女雇用機会均等法施行後の三時点比較―
		税負担と労働供給 －「日本労働研究雑誌」No.605 2010.12 －
		若者の有業・無業状態における属性の考察
		女性の働き方と少子化に関する考察
高等教育目的	就業構造基本調査 全国消費実態調査 社会生活基本調査	一橋大学大学院経済学研究科「演習」（労働経済学Ⅰ）

論文等：http://www.nstac.go.jp/services/jisseki.html

（参考資料３）オーダーメード集計の利用実績例

調査名	研究の名称
国　勢　調　査	岡山大学大学院社会文化科学研究科紀要第35号（2013.3） ・2005年国勢調査にみる在日外国人女性の結婚と仕事・住居 ・2005年国勢調査にみる外国人の教育 ・2005年国勢調査にみる在日外国人の仕事
	Access to childcare and the employment of women with preschool-aged children in Tokyo
	看護人材の就業率の推移 －再検討した潜在者数推計方法による結果から－
	地域別経済指標に基づくSDモデルの開発
	在日外国人の仕事 －2000年国勢調査データの分析から－
	家族・ジェンダーからみる在日外国人 －2000年国勢調査データの分析から－
	在学率と通学率から見る在日外国人青少年の教育 －2000年国勢調査データの分析から－
	エリア・サンプリングにおける問題点
全国消費実態調査	新・家計消費論 －高齢層が支える都市部消費－
就業構造基本調査	近年における都道府県別貧困率の推移について ―ワーキングプアを中心に

論文等：http://www.nstac.go.jp/services/jisseki.html

参考資料4：公的統計部局の課題：提供データ範囲の拡大：
統計センターにおける匿名データ及びオーダーメード集計の提供

提供データ一覧（厚生労働省が別途国民生活基礎調査提供）

調査名（年次等）	匿名	オーダー
国勢調査		
昭和55年		○
昭和60年		○
平成2年		○
平成7年		○
平成12年	○	○
平成17年	○	○
平成22年		○
就業構造基本調査		
昭和54年		○
昭和57年		○
昭和62年		○
平成4年	○	○
平成9年	○	○
平成14年	○	○
平成19年		○
平成24年		○

調査名（年次等）	匿名	オーダー
住宅・土地統計調査		
昭和53年		○
昭和58年		○
昭和63年		○
平成5年	○	○
平成10年	○	○
平成15年	○	○
平成20年		○
平成25年		●
全国消費実態調査		
平成元年	○	
平成6年	○	
平成11年	○	●
平成16年	○	○
平成21年		○

調査名（年次等）	匿名	オーダー
社会生活基本調査【調査票A】生活時間編・生活行動編		
昭和56年		○
昭和61年		○
平成3年	○	○
平成8年	○	○
平成13年	○	○
平成18年	○	○
平成23年		○
社会生活基本調査【調査票B】生活時間編		
平成13年	○	
平成18年	○	
労働力調査【基礎調査票】		
昭和55年1月～63年12月		○
平成元年1月～23年12月	●	○
平成24年1月～26年12月		○
労働力調査【特定調査票】		
平成14年1月～26年12月		○

調査名（年次等）	匿名	オーダー
家計調査		
昭和56年1月～平成26年12月		●
家計消費状況調査		
平成14年1月～26年12月		○
消費動向調査		
平成16年4月～27年3月		○
企業行動に関するアンケート調査		
平成18年度～26年度		○
賃金構造基本統計調査		
平成18年～26年		○
学校基本調査【高等教育機関編】		
平成20年度～26年度		●
学校基本調査【初等教育機関編】		
平成20年度～22年度		○
建設着工統計調査		
平成21年4月～27年3月		○

○印は提供中、●印は（複数年存在するものについては最新年（度）分を）平成27年度中に提供予定

～　利用可能な統計調査は、今後、さらに拡大予定　～

科研プロジェクト「経済統計・政府統計の理論と応用」
2016年 1月29日（金）
東京大学経済学部小島ホール

地方創生と人口統計

― まち・ひと・しごと再生の課題 ―

国立社会保障・人口問題研究所
金子隆一

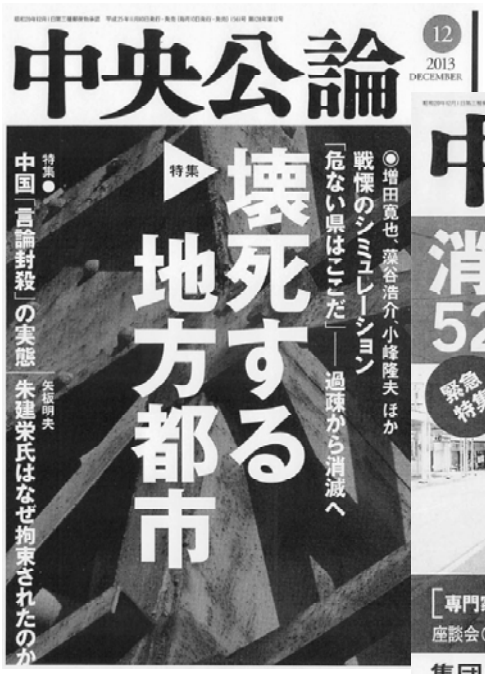
話しのアウトライン

1．経　緯
　日本創成会議～創生本部～自治体

2．地域人口の概観
　地方消滅 と 東京一極集中
　“少子化対策”と人口１億人維持

3．統計データの利用
　RESAS

（１）日本創生会議 増田寛也＋人口減少問題検討分科会



(2013.12)

2040年に若年女性が2010年の50%以下に減少する自治体（＝“消滅可能性都市”）が半数に昇る！

(2014.6)

(2014.8)

(2014.7)

（２）政府の動き

社会保障改革国民会議

社会保障改革国民会議 (2013.8)

報告書ポイント

社会保障制度改革国民会議 委員名簿
2013年4月19日現在（◎は会長、○は会長代理）

伊藤元重	東京大学大学院経済学研究科教授
○ 遠藤久夫	学習院大学経済学部長
大島伸一	国立長寿医療研究センター総長
大日向雅美	恵泉女学園大学大学院平和学研究科教授
権丈善一	慶應義塾大学商学部教授
駒村康平	慶應義塾大学経済学部教授
榊原智子	読売新聞東京本社編集局社会保障部次長
神野直彦	東京大学名誉教授
◎ 清家 篤	慶應義塾長
永井良三	自治医科大学学長
西沢和彦	日本総合研究所調査部上席主任研究員
増田寛也	野村総合研究所顧問
宮武 剛	目白大学大学院生涯福祉研究科客員教授
宮本太郎	中央大学法学部教授
山崎泰彦	神奈川県立保健福祉大学名誉教授

（1）「1970 年代モデル」から「21 世紀（2025 年）日本モデル」へ
（2）すべての世代を対象とし、すべての世代が相互に支え合う仕組み
（3）女性、若者、高齢者、障害者などすべての人々が働き続けられる社会
（4）すべての世代の夢や希望につながる子ども・子育て支援の充実
（5）低所得者・不安定雇用の労働者への対応
（6）地域づくりとしての医療・介護・福祉・子育て
（7）国と地方が協働して支える社会保障制度改革
（8）成熟社会の構築へのチャレンジ （世界の先頭を歩む）

・人口構造（の変化）を前提とした社会保障の重要性の認識
・社会保障の枠の拡大 → 少子化への対策等を（明確に）組み入れる
・地域への着目

選択する未来

未来への選択 － 人口急減・超高齢社会を超えて、日本発成長・発展モデルを構築 －

【（確実な未来）】
50年後も、人口減少が続き、加速。
現状のままでは、「人口急減・超高齢社会」の到来

人口 （2013年）12,730万人 → （2060年）8,674万人
減少幅
- 2008～2013年 ▲16万人／年
- 2010年代後半～2020年代初頭 ▲50～60万人／年
- 2040年代初頭 ▲100万人／年

高齢化率 （2013年）25% → （2060年）40%

【現状のまま何もしない場合の未来像】
・プラス成長を続けることは困難になり、マイナス成長が定着
・「人口オーナス」と「縮小スパイラル」の双方が作用し、国民生活低下のおそれ
・女性、高齢者、若者が活躍できない労働市場の二極化、格差の固定化・再生産
・地方で4分の1以上の自治体が消滅可能性、東京では超高齢化
・医療・介護費の増加により財政破たんリスクの高まり

危機意識の共有 → 未来を変える時間軸
- 成長・発展（早期に）：アベノミクスを機に長期の発展経路へ
- 人口・地域（2020年目途）：人口減少が加速する前にトレンドを変える
- 2020年のその先へ：持続的・安定的に成長・発展する経済社会

【未来への選択】
・制度、政策や人々の意識が速やかに変われば、「人口急減・超高齢社会」への流れは変えられる
・若い世代や次の世代が豊かさを得て、結婚し、子どもを産み育てることができるよう集中して改革・変革

①人口
50年後に1億人程度（この場合、その一世代後には微増に転じる）
- 国民の希望どおり子どもを産み育てられる環境により、1億人程度の人口を保持
- 資源配分を高齢者から子どもへシフト、出産・子育て支援を倍増。費用は現世代で負担
- 子どものための政策推進

②成長・発展
経済を世界に開き、「創意工夫による新たな価値の創造」により、成長し続ける
- イノベーションが生産性向上の切り札
- 産業・企業の「新陳代謝・若返り」（ダイナミズム）
- オープンな国づくりと、外国人材の戦略的受け入れ
- 債務残高対GDP比引下げ等の明確な目標

③人の活躍
年齢、性別に関わらず能力発揮
- 男女の働き方改革により、能力や意欲に応じた活躍の機会充実
- 70歳まで働ける社会（新生産年齢人口）
- 未来の技術や産業に適応したプレイヤーの育成
- 格差の再生産の回避

④地域の未来
個性を活かした地域戦略、集約・活性化
- 新しい発想で資源を利用し、働く場所をつくる（農業、観光等）
- 「集約・活性化」によるコンパクトな地域・地方中枢都市圏域の形成
- 東京への若者の人口流出を抑制
- 東日本大震災の復興を地域のモデルに

⑤信頼・規範
基盤的な制度、文化、公共心など社会の土台を大切にする
- 日本の国土に育まれた伝統、文化、美意識、価値観の継承・発信
- 国際貢献やルールづくりへ参加、世界に発信し続ける
- 社会保障制度や財政の持続可能性の確保

骨太の方針

○経済財政運営と改革の基本方針2014 ～デフレから好循環拡大へ～
（平成26 年6月24 日閣議決定）（抜粋）

(2014.6)

第１章　アベノミクスのこれまでの成果と今後の日本経済の課題

４．日本の未来像に関わる制度・システムの改革
（「人口急減・超高齢化」 の克服）

デフレ脱却・経済再生の先に、もう一つ超えなければならない高いハードルがある。現在の日本は、「人口急減・超高齢化」へ確実に向かっている。この流れを変えなければ、持続的・安定的な成長軌道に乗っていくことはできない。

人口急減・超高齢化の流れを変えることは容易でなく、流れが変わっても効果が現れるまで長期間を要する。人口急減・超高齢化の流れを変えられない場合には、経済規模が収縮し、縮小スパイラルに陥るおそれがある。そこに至っては、もはや回復は困難となろう。従来の少子化対策の枠組みにとらわれず、福祉分野以外にも、教育、社会保障、社会資本整備、地方行財政、産業振興、税制など、あらゆる分野の制度・システムを若者・子ども世代や次の世代のためになっているか、結婚しやすく子育てしやすい環境を実現する仕組みになっているかという観点から見直し、2020年を目途にトレンドを変えるために抜本的な改革・変革を推進すべき時期に来ている。

希望通りに働き、結婚、出産、子育てを実現することができる環境を整え、人々の意識が大きく変わり、2020年を目途にトレンドを変えていくことで、50年後にも１億人程度の安定的な人口構造を保持することができると見込まれる。

まち・ひと・しごと創生本部の組織体制 (2014.9)

創生本部

まち・ひと・しごと創生本部

○設置根拠：閣議決定
○構成：
本部長　総理大臣
副本部長　地方創生担当大臣、官房長官
本部員　他の全ての国務大臣
※その他必要に応じて本部長が出席を求める

まち・ひと・しごと創生会議 民間有識者（12 名）
池田 弘　公益社団法人日本ニュービジネス協議会連合会会長
伊東 香織　岡山県倉敷市長
おおこや
大社 充　NPO 法人グローバルキャンパス理事長
奥田 麻依子　島根県海士町 隠岐島前高校魅力化コーディネーター
坂根 正弘　コマツ相談役
清水 志摩子　NPO 法人全国商店街おかみさん会理事長
田中 進　農業生産法人(株)サラダボウル代表取締役
冨山 和彦　経営共創基盤代表取締役CEO
中橋 恵美子　NPO 法人わはねネット理事長
樋口 美雄　慶應義塾大学商学部教授
増田 寛也　東京大学公共政策大学院客員教授
山本 眞樹夫　帯広畜産大学監事、前小樽商科大学長

まち・ひと・しごと創生会議

○設置根拠：本部長決定
○構成：
議長　総理大臣
副議長　地方創生担当大臣、官房長官
議員　・経済財政担当大臣、少子化担当大臣、復興大臣、総務大臣、財務大臣、文部科学大臣、厚生労働大臣、農林水産大臣、経済産業大臣、国土交通大臣
・民間有識者
※その他必要に応じて議長が出席を求める

まち・ひと・しごと創生本部事務局

○設置根拠：総理決定
○構成：
事務局長　官房副長官（事務）
事務局長代行　・総理大臣補佐官（地方創生等担当）
・官房副長官補（内政）
事務局長代理　（3名）

まち・ひと・しごと創生本部幹事会

○設置根拠：本部長決定
○構成：
議長　地方創生担当大臣
議長代理　地方創生担当副大臣、官房副長官（事務）
副議長　地方創生担当大臣政務官、総理大臣補佐官、官房副長官補
議員　事務局長代理、全事務次官・長官

まち・ひと・しごと創生法案の概要 (2014.11)

創生法

少子高齢化の進展に的確に対応し、人口の減少に歯止めをかけるとともに、東京圏への人口の過度の集中を是正し、それぞれの地域で住みよい環境を確保して、将来にわたって活力ある日本社会を維持していくために、まち・ひと・しごと創生（※）に関する施策を総合的かつ計画的に実施する。

※まち・ひと・しごと創生：以下を一体的に推進すること。
- まち…国民一人一人が夢や希望を持ち、潤いのある豊かな生活を安心して営める地域社会の形成
- ひと…地域社会を担う個性豊かで多様な人材の確保
- しごと…地域における魅力ある多様な就業の機会の創出

基本理念（第２条）

①国民が個性豊かで魅力ある地域社会で潤いのある豊かな生活を営めるよう、それぞれの地域の実情に応じた環境を整備

②日常生活・社会生活の基盤となるサービスについて、需要・供給を長期的に見通しつつ、住民負担の程度を考慮して、事業者・住民の理解・協力を得ながら、現在・将来における提供を確保

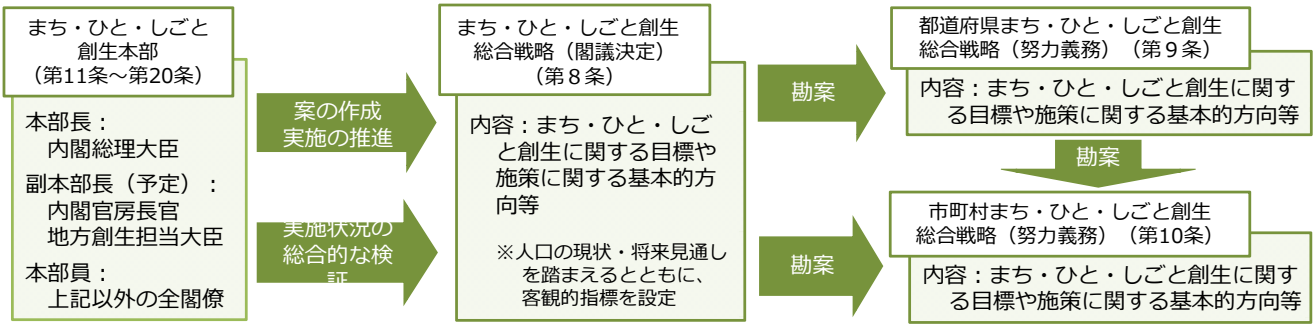
③結婚・出産は個人の決定に基づくものであることを基本としつつ、結婚・出産・育児について希望を持てる社会が形成されるよう環境を整備

④仕事と生活の調和を図れるよう環境を整備

⑤地域の特性を生かした創業の促進・事業活動の活性化により、魅力ある就業の機会を創出

⑥地域の実情に応じ、地方公共団体相互の連携協力による効率的かつ効果的な行政運営の確保を図る

⑦国・地方公共団体・事業者が相互に連携を図りながら協力するよう努める



施行期日：公布日（創生本部・総合戦略に関する規定は、公布日から１か月を超えない範囲内で政令で定める日）

まち・ひと・しごと創生「長期ビジョン」

長期ビジョン

３つの視点

- 若い世代の就労・結婚・子育ての希望の実現
- 『東京一極集中』の歯止め
- 地域の特性に即した地域課題の解決

「しごと」と「ひと」の好循環を実現するための、４つの目標

①地方における安定的な雇用を創出
②地方への新しいひとの流れをつくる
③若い世代の結婚・出産・子育ての希望をかなえる
④時代に合った地域をつくり、安心な暮らしを守るとともに、地域と地域を連携する

魅力あふれる地方を創生

まち・ひと・しごと創生「長期ビジョン」と「総合戦略」の全体像

資料４－２ 創生

長期ビジョン

中長期展望（2060 年を視野）

Ⅰ．人口減少問題の克服

◎2060 年に1億人程度の人口を維持

◆人口減少の歯止め
・国民の希望が実現した場合の出生率（国民希望出生率）＝1.8

◆「東京一極集中」の是正

Ⅱ．成長力の確保

◎2050 年代に実質 GDP 成長率 1.5～2%程度維持

（人口安定化、生産性向上が実現した場合）

総合戦略（2015～2019 年度の5か年）

基本目標（成果指標、2020 年）	主な重要業績評価指標（KPI）（※1）	主な施策
「しごと」と「ひと」の好循環作り		
地方における安定した雇用を創出する ◆若者雇用創出数（地方） 2020 年までの5年間で 30 万人 ◆若い世代の正規雇用労働者等の割合 2020 年までに全ての世代と同水準 （15～34 歳の割合：92.2%（2013 年） （全ての世代の割合：93.4%（2013 年） ◆女性の就業率 2020 年までに 73% （2013 年 70.8%）	農林水産業の成長産業化 6次産業市場 10 兆円：就業者数5万人創出 訪日外国人旅行消費3兆円へ（2013 年 1.4 兆円）：雇用数8万人創出 地域の中核企業、中核企業候補 1,000 社支援：雇用数8万人創出	①地域産業の競争力強化（業種横断的取組） ・包括的創業支援、中核企業支援、地域イノベーション推進、対内直投促進、金融支援 ②地域産業の競争力強化（分野別取組） ・サービス産業の付加価値向上、農林水産業の成長産業化、観光、ローカル版クールジャパン、ふるさと名物、文化・芸術・スポーツ ③地方への人材還流、地方での人材育成、雇用対策 ・「地域しごと支援センター」の整備・稼働 ・「プロフェッショナル人材センター」の稼働
地方への新しいひとの流れをつくる 現状：東京圏年間 10 万人入超 ◆地方・東京圏の転出入均衡（2020 年） ・地方→東京圏転入　6万人減 ・東京圏→地方転出　4万人増	地方移住の推進 ：年間移住あっせん件数 11,000 件 企業の地方拠点機能強化 ：拠点強化件数 7,500 件、雇用者数4万人増加 地方大学活性化：自県大学進学者割合平均 36%（2013 年度 32.9%）	① 地方移住の推進 ・「全国移住促進センター」の開設、移住情報一元提供システム整備 ・「地方居住推進国民会議」（地方居住（二地域居住を含む）推進） ・「日本版 CCRC※2」の検討、普及 ②地方拠点機能強化、地方採用・就労拡大 ・企業の地方拠点強化等 ・政府関係機関の地方移転 ・遠隔勤務（サテライトオフィス、テレワーク）の促進 ③地方大学等創生5か年戦略
若い世代の結婚・出産・子育ての希望をかなえる ◆安心して結婚・妊娠・出産・子育てできる社会を達成していると考える人の割合 40%以上（2013 年度 19.4%） ◆第１子出産前後の女性継続就業率 55%（2010 年 38%） ◆結婚希望実現指標 80%（2010 年 68%） ◆夫婦子ども数予定（2.12）実績指標 95%（2010 年 93%）	若い世代の経済的安定：若者就業率 78% （2013 年 75.4%） 妊娠・出産・子育ての切れ目のない支援 ：支援ニーズの高い妊産婦への支援実施 100% ワーク・ライフ・バランス実現：男性の育児休業取得率 13%（2013 年 2.03%）	①若者雇用対策の推進、正社員実現加速 ②結婚・出産・子育て支援 ・「子育て世代包括支援センター」の整備 ・子ども・子育て支援新制度の円滑かつ持続的な実施 ・多子世帯支援、三世代同居・近居支援 ③仕事と生活の調和（ワーク・ライフ・バランス）の実現（働き方改革） ・育児休業の取得促進、長時間労働の抑制、企業の取組の支援等
好循環を支える、まちの活性化		
時代に合った地域をつくり、安心な暮らしを守るとともに、地域と地域を連携する ◆地域連携数など ※目標数値は地方版総合戦略を踏まえ設定	「小さな拠点」の形成 ：「小さな拠点」形成数 定住自立圏の形成 ：協定締結等圏域数（140 圏域） 既存ストックのマネジメント ：中古・リフォーム市場規模 20 兆円 （2010 年 10 兆円）	①「小さな拠点」（多世代交流・多機能型）の形成支援 ②地方都市における経済・生活圏の形成（地域連携） ・都市のコンパクト化と周辺等のネットワーク形成 ・「連携中枢都市圏」の形成、・定住自立圏の形成促進 ③大都市圏における安心な暮らしの確保 ④既存ストックのマネジメント強化

※1　Key Performance Indicator の略。政策ごとの達成すべき成果目標として、日本再興戦略（2013 年６月）でも設定されている。

※2　米国では高齢者が移り住み、健康時から介護・医療が必要となる時期まで継続的なケアや生活支援サービス等を受けながら生涯学習や社会活動等に参加するような共同体（Continuing Care Retirement Community）が約 2,000 カ所ある。

地方への多様な支援と「切れ目」のない施策の展開

資料４ 創生

国

国の長期ビジョン：2060 年に1億人程度の人口を維持する中長期展望を提示

国の総合戦略：2015～2019 年度（5か年）の政策目標・施策を策定

地方

地方人口ビジョン：各地域の人口動向や将来人口推計の分析や中長期の将来展望を提示

地方版総合戦略：各地域の人口動向や産業実態等を踏まえ、2015～2019 年度（5か年）の政策目標・施策を策定

情報支援

○「地域経済分析システム」

・各地域が、産業・人口・社会インフラなどに関し必要なデータ分析を行い、各地域に即した地域課題を抽出し対処できるよう、国は「地域経済分析システム」を整備。

＜地方自治体の戦略策定と国の支援＞

・地方が自立につながるよう自らが考え、責任を持って戦略を推進。
・国は「情報支援」、「人的支援」、「財政支援」を切れ目なく展開。

人的支援

○「地方創生人材支援制度」

・小規模市町村に国家公務員等を首長の補佐役として派遣。

○「地方創生コンシェルジュ制度」

・市町村等の要望に応じ、当該地域に愛着・関心を持つ、意欲ある府省庁の職員を相談窓口として選任。

財政支援

○「地方版総合戦略」の策定・実施の財政的支援

緊急的取組

経済対策（まち・ひと・しごと創生関連）

○地域住民生活等緊急支援のための交付金（仮称）

地方創生先行型の創設

地方の積極的な取組を支援する自由度の高い交付金を、26 年度補正予算で先行的に創設。地方版総合戦略の早期かつ有効な策定・実施には手厚く支援。対象事業は、①地方版総合戦略の策定、②地方版総合戦略における「しごとづくりなど」の事業。メニュー例：UIJターン助成、創業支援、海外販路開拓など。

地域消費喚起・生活支援型

メニュー例：
プレミアム付商品券
低所得者等向け灯油等購入助成
ふるさと名物商品・旅行券　　等

27年度

総合戦略に基づく取組

○国：27年度を初年度とする「総合戦略」を推進。
○地方：国の総合戦略等を勘案し、「地方人口ビジョン」及び「地方版総合戦略」を策定し、施策を推進。

税制・地方財政措置

○企業の地方拠点強化に関する取組を促進するための税制措置

○地方創生の取組みに要する経費について地方財政計画に計上し、地方交付税を含む地方の一般財源確保　　等

28年度以降

総合戦略に基づく取組

○総合戦略の更なる進展

新型交付金の本格実施へ

○地方版総合戦略に基づく事業・施策を自由に行う
○客観的な指標の設定・PDCA による効果検証を行う

2．地域人口の概観

地方消滅 と 東京一極集中

“少子化対策”と人口１億人維持

■ 自治体の「消滅可能性」について

【日本創成会議（人口減少問題検討分科会）】
今後の「人口移動が収束しない」とすると、2040年に若年女性が2010年の50%以下に減少する自治体（＝“消滅可能性都市”）は 896（全体の 49.8%）となり、さらにその中で人口規模が１万人を下回る自治体（＝“消滅可能性が高い”自治体）は 523（全体の 29.1%）となる。

● 社人研推計（平成24年３月推計）では、2005～10年の純移動率は2015～20年に１／２となることを基本仮定としているが（自治体の類型化による差異あり）、これによると 2040年の"消滅可能性都市"は、373（全体の20.７%）、"消滅可能性が高い"自治体は、243 (13.5%）である（図表）。

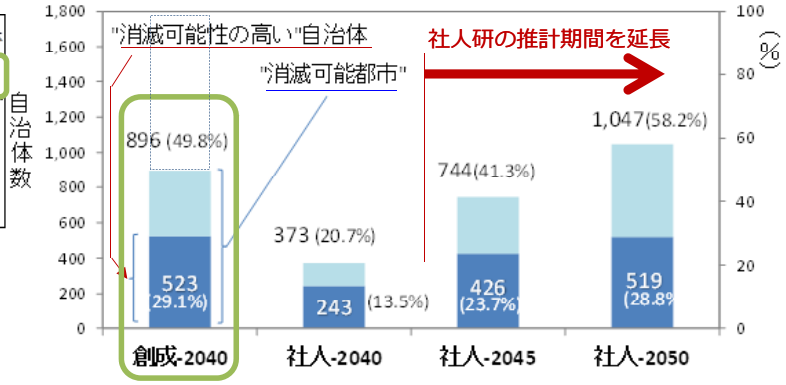
● しかし、社人研推計においても約10年の差で、創成会議推計とほぼ同様の状況が現出する（図表）。

● すなわち、創成会議推計は社人研推計と比較して、人口減少、年齢構造変化（高齢化）についての変化ペースが速いだけで、長期的な帰結に違いはないと見ることができる。

創成会議推計と社人研推計の比較

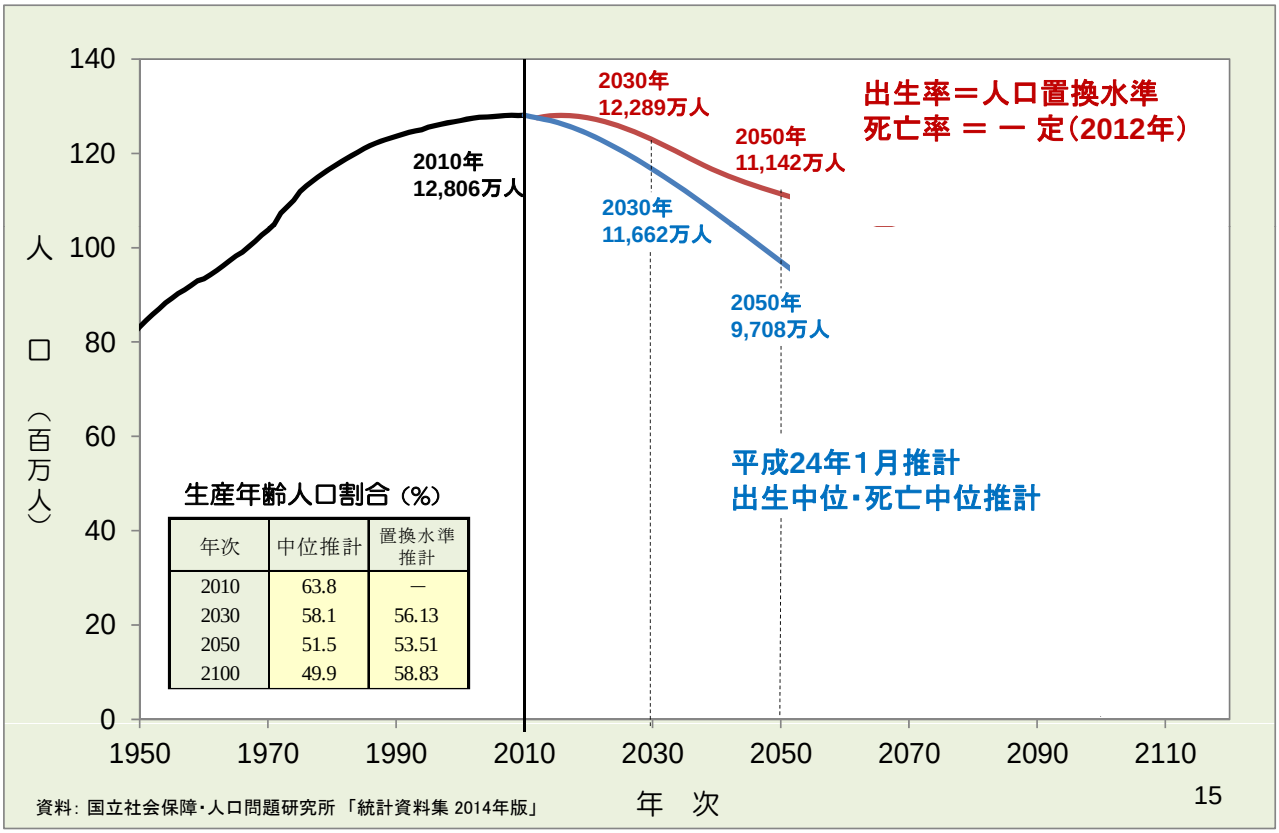
	"消滅可能性都市"	"消滅可能性の高い"自治体
創成-2040	896 (49.8%)	523 (29.1%)
社人-2040	373 (20.7%)	243 (13.5%)
社人-2045	744 (41.3%)	426 (23.7%)
社人-2050	1,047 (58.2%)	519 (28.8%)

※ 自治体総数（2010年） 1,800



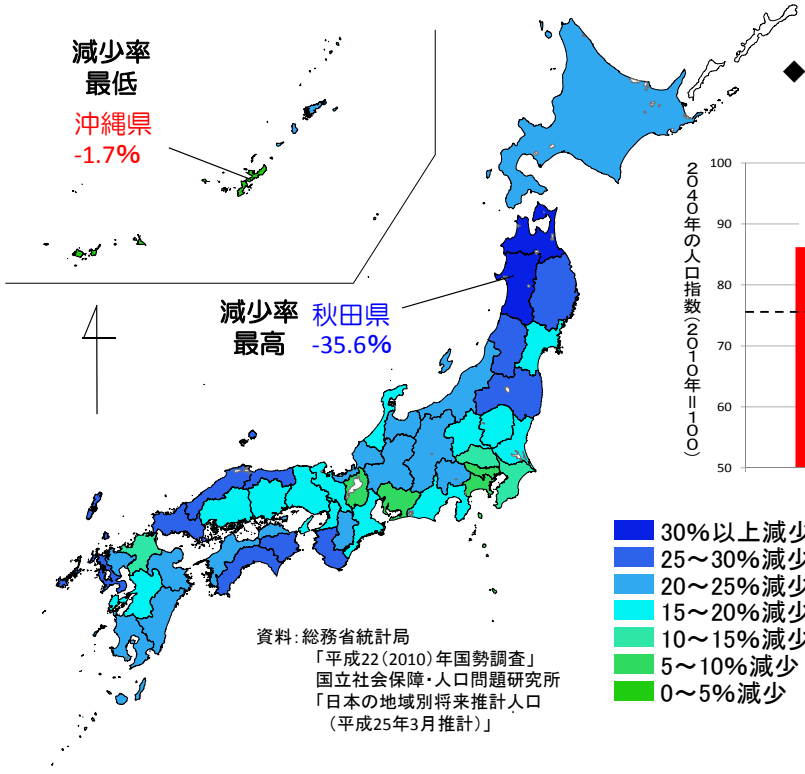
※ 社人－2045,2050は、2035～40年の移動率を含めた動態率仮定値を固定して延長推計試算したもの（公表値ではない）.

人口減少に対する人口モメンタムの効果



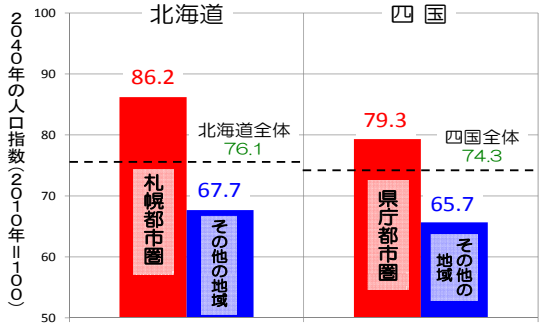
人口減少の程度は、地域差が大きい

都道府県別にみた 2010年 → 2040年の人口増減率

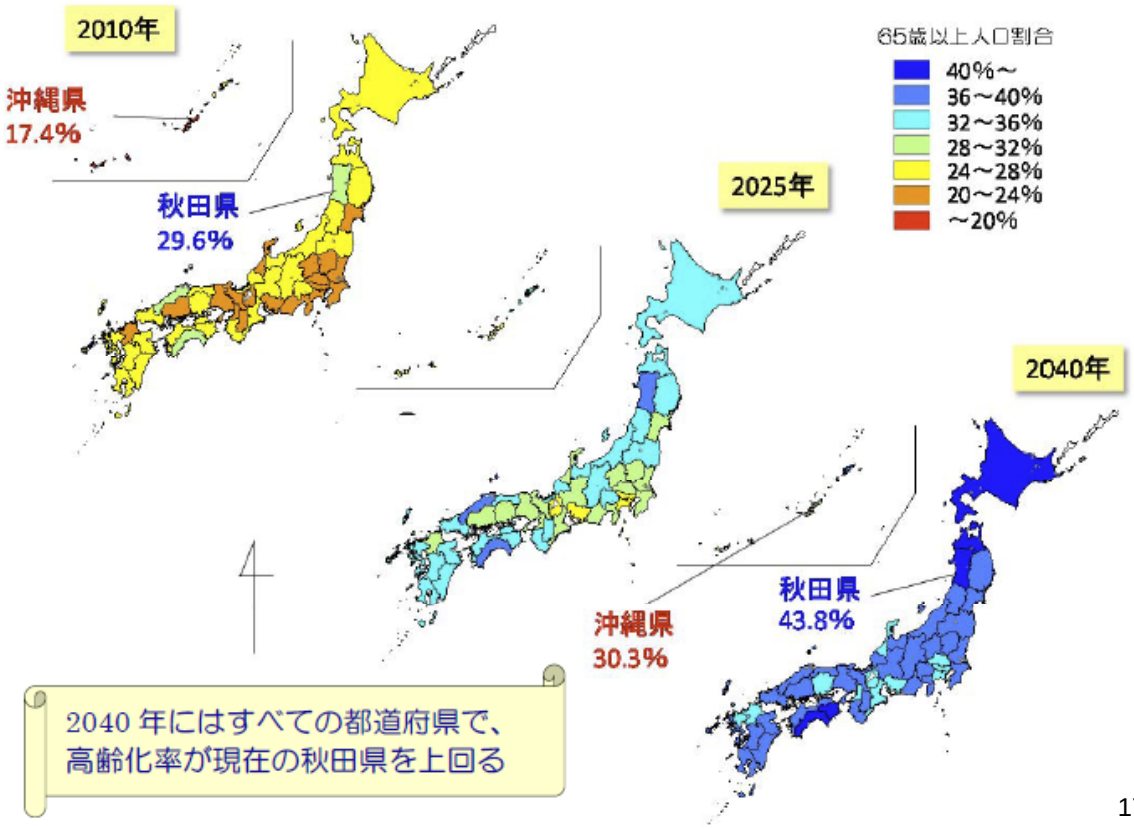


◆ 今後、東北、山陰、四国などでの人口減少が著しい

◆ ただし、同一都道府県内でも、都市部と農漁村地域では人口動向は大きく異なる

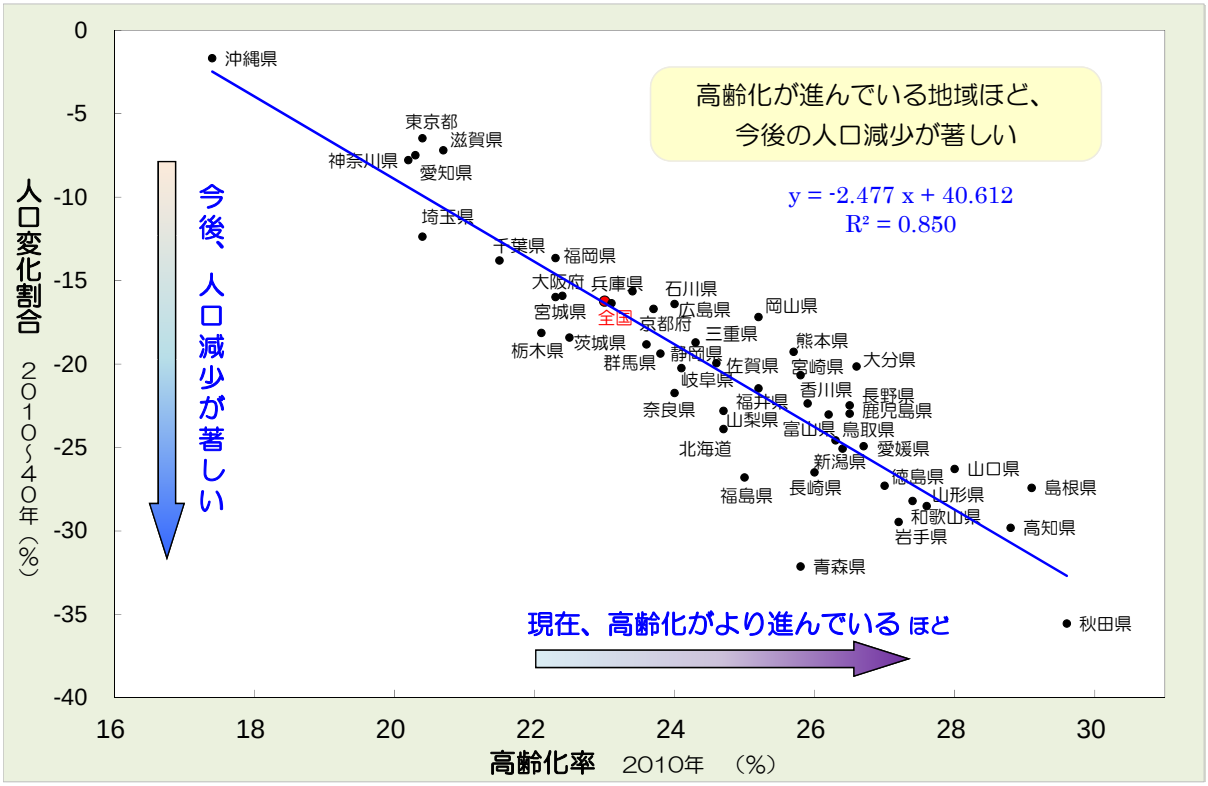


都道府県別にみた高齢化率の増加：2010, 25, 40年



2040 年にはすべての都道府県で、
高齢化率が現在の秋田県を上回る

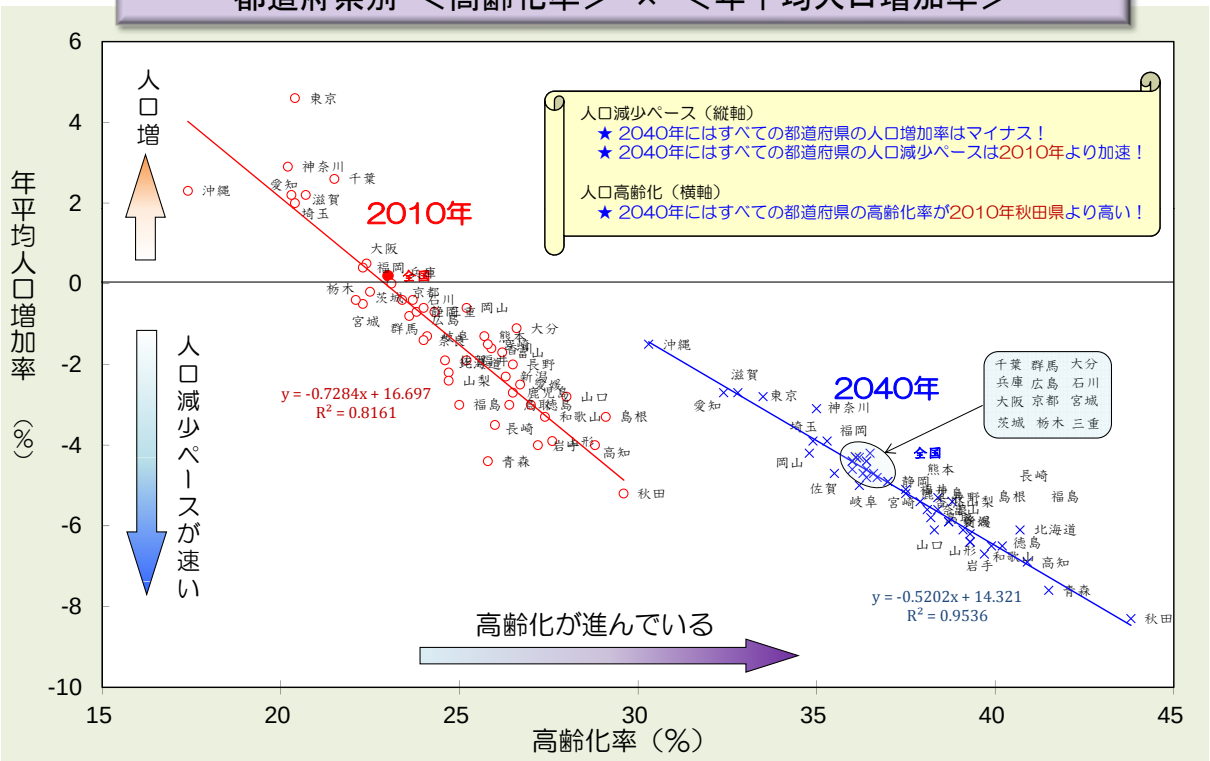
高齢化が進んでいる地域では、今後の人口の減少が著しい



資料：総務省統計局「国勢調査」、国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

高齢化が進んでいる地域ほど、人口減少のペースが早い

都道府県別 ＜高齢化率＞ × ＜年平均人口増加率＞



資料：総務省統計局「国勢調査」、国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

１－２　日本の将来人口動向

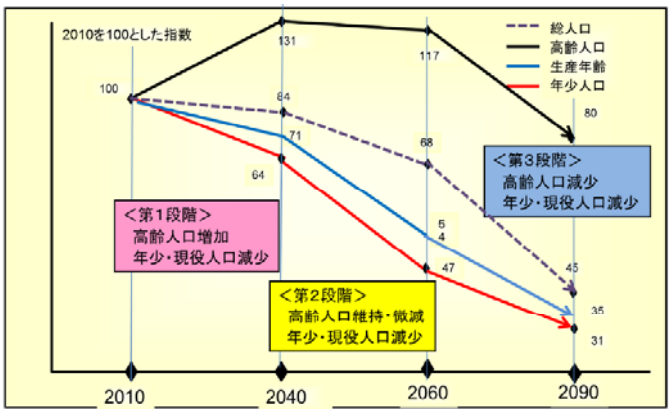
創生

○　人口減少は世代別の異なる動きの中で進む。

○　日本の将来人口動向は、第1段階：高齢人口が増加する時期、第2段階：高齢人口が維持・微減となる時期、第3段階：高齢人口さえも減少する時期、に大きく分けられる。

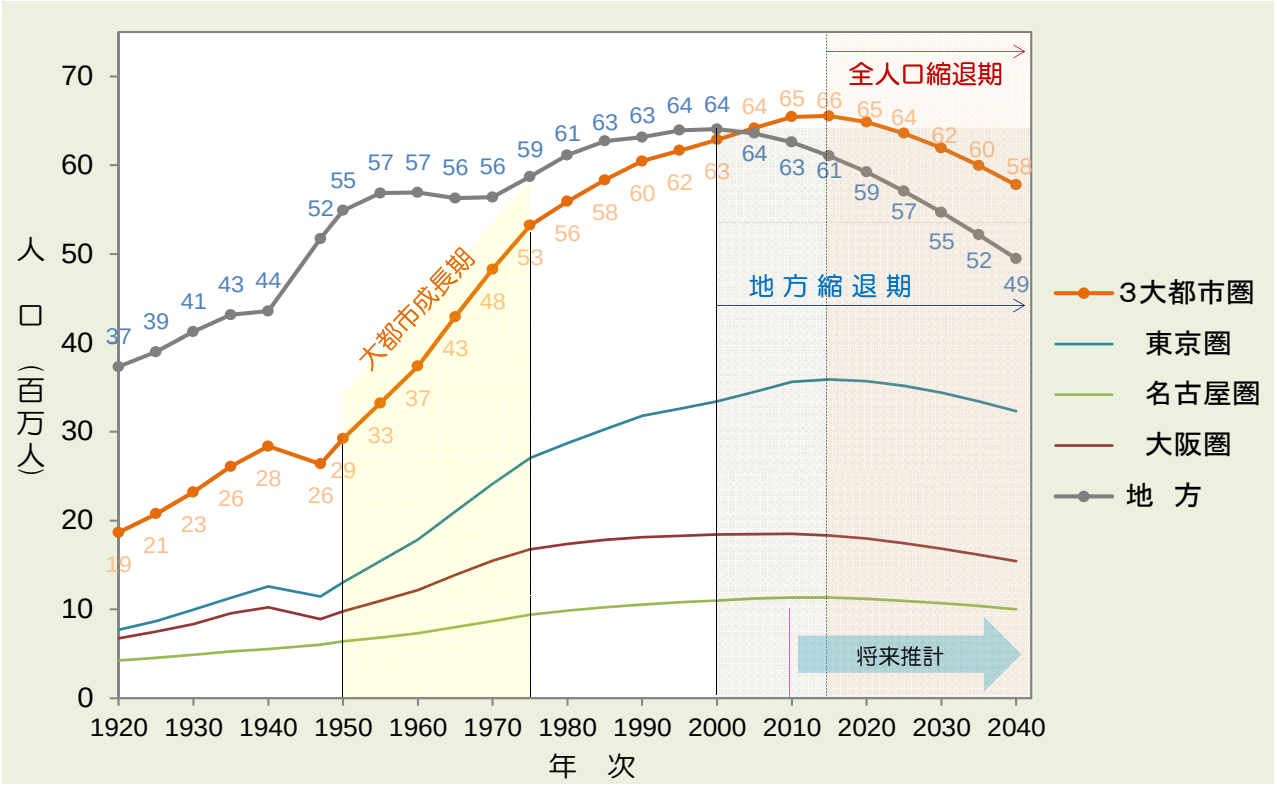
将来推計人口【中位推計−合計特殊出生率1.35】

	2010年	2040年	2060年	2090年	2110年
総人口	12,806万人	10,728万人	8,674万人	5,727万人	4,286万人
老年人口（65歳以上） 高齢化率	2,948万人 23.0%	3,878万人 36.1%	3,464万人 39.9%	2,357万人 41.2%	1,770万人 41.3%
生産年齢人口（15～64歳）	8,174万人	5,787万人	4,418万人	2,854万人	2,126万人
年少人口（～14歳）	1,684万人	1,073万人	792万人	516万人	391万人



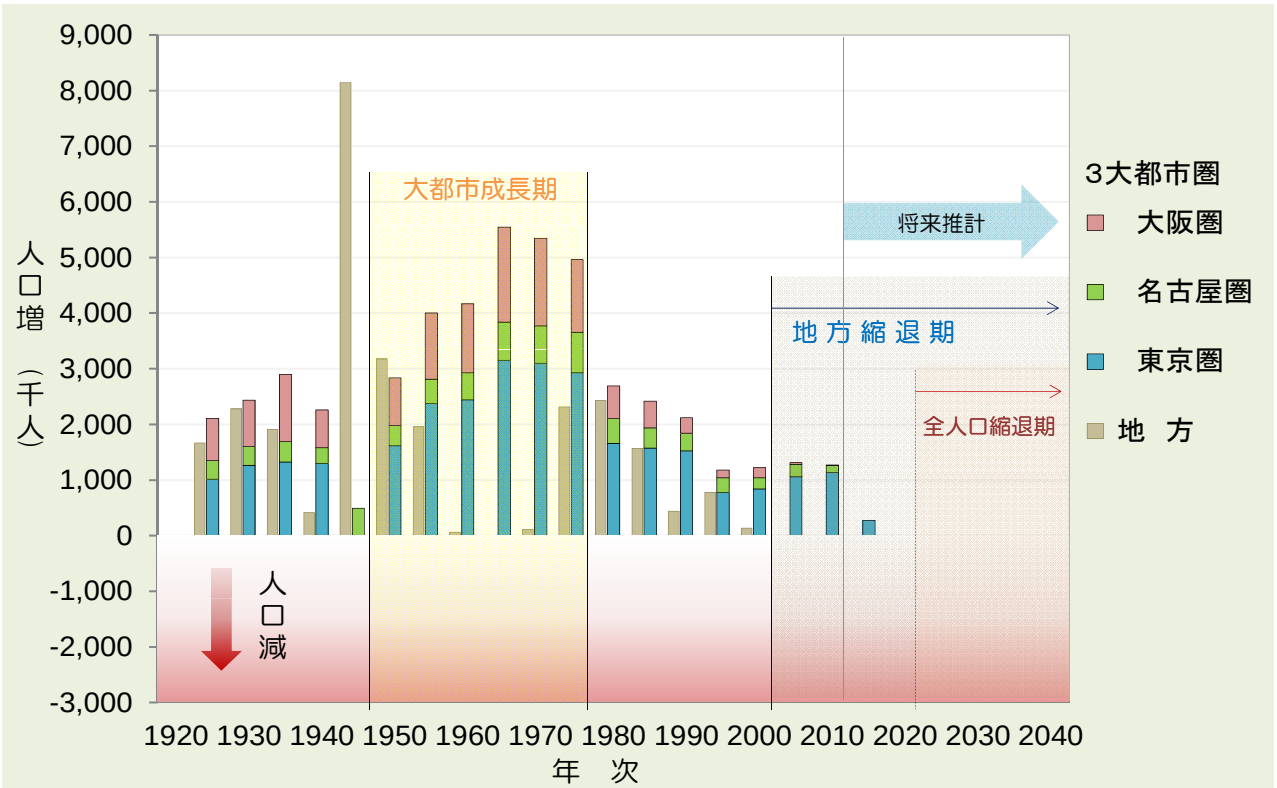
（備考）国立社会保障・人口問題研究所「日本の将来推計人口（平成24年1月推計）」より作成

一極集中？ 都市圏と地方－人口の推移



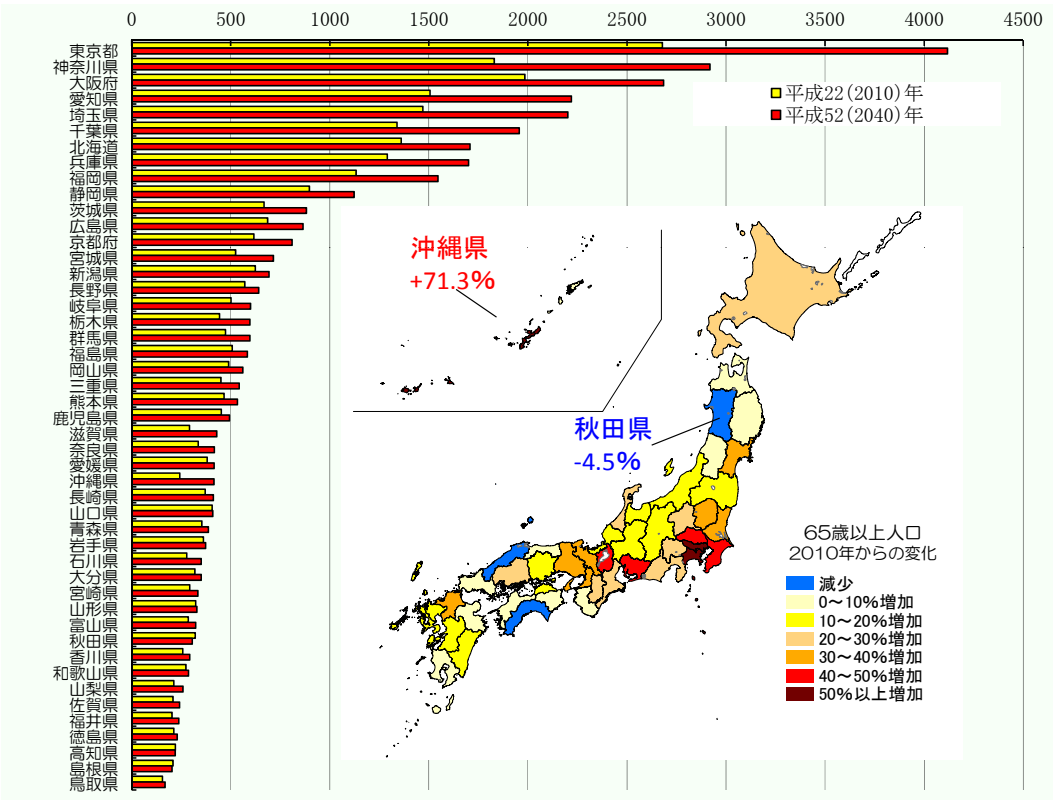
資料： 2010年以前：総務省統計局「国勢調査」、2015年以降：国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

都市圏と地方－人口増減の推移（５年間隔）



資料： 2010年以前：総務省統計局「国勢調査」、2015年以降：国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

都道府県別,６５歳以上人口の変化 ：2010年、2040年比較



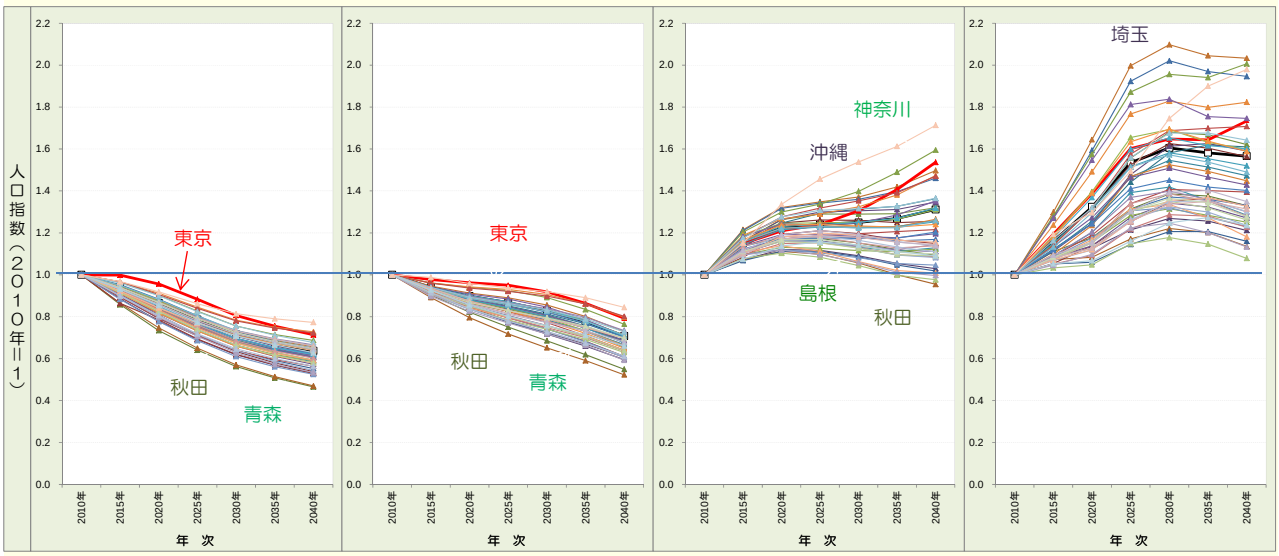
資料：総務省統計局「国勢調査」、国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

年齢階層による人口増減の違い：都道府県別

★ わが国の人口は、今後しばらく年齢階層（年少、生産年齢、高齢）ごとに大きく異なる推移をする。
→ 年齢が若いほど減少が著しく、高齢になるほど増加が著しい。

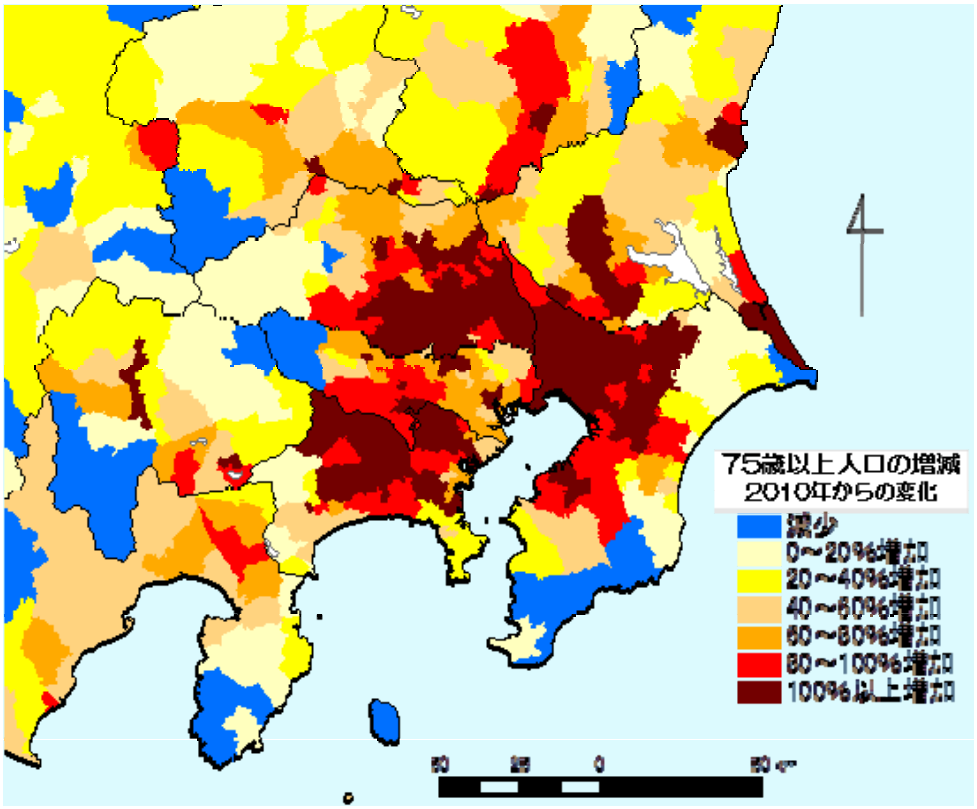
★ 相対的変化には、地域（都道府県）による違いが有るが、その違いは高齢ほど著しい。

都道府県別にみた 年齢階層別 人口推移 （2010年人口＝1.0）



資料：国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」.

首都圏の高齢化 ： 2010→2040年 75歳以上 人口増減率



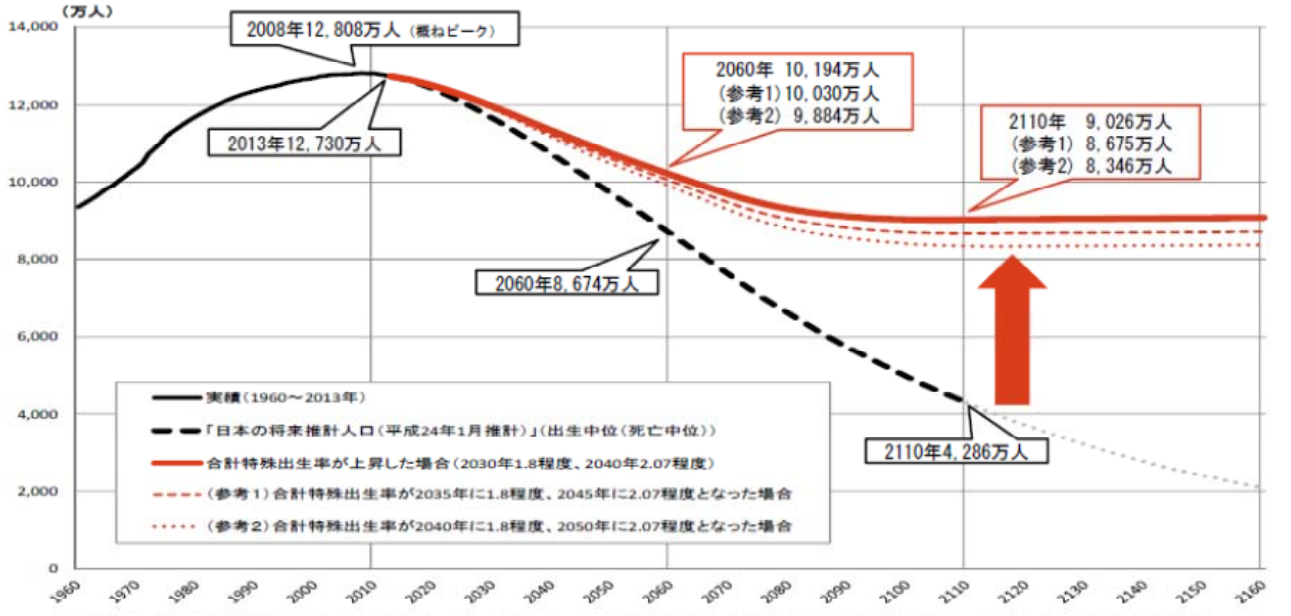
資料：総務省統計局「国勢調査」、国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」

一億人？

図１． 我が国の人口の推移と長期的な見通し

創生

○ 国立社会保障・人口問題研究所「日本の将来推計人口（平成24年1月推計）」（出生中位（死亡中位））によると、2060年の総人口は約8,700万人まで減少すると見通されている。
○ 仮に、合計特殊出生率が2030年に1.8程度、2040年に2.07程度（2020年には1.6程度）まで上昇すると、2060年の人口は約1億200万人となり、長期的には9,000万人程度で概ね安定的に推移するものと推計される。
○ なお、仮に、合計特殊出生率が1.8や2.07となる年次が５年ずつ遅くなると、将来の定常人口が概ね300万人程度少なくなると推計される。



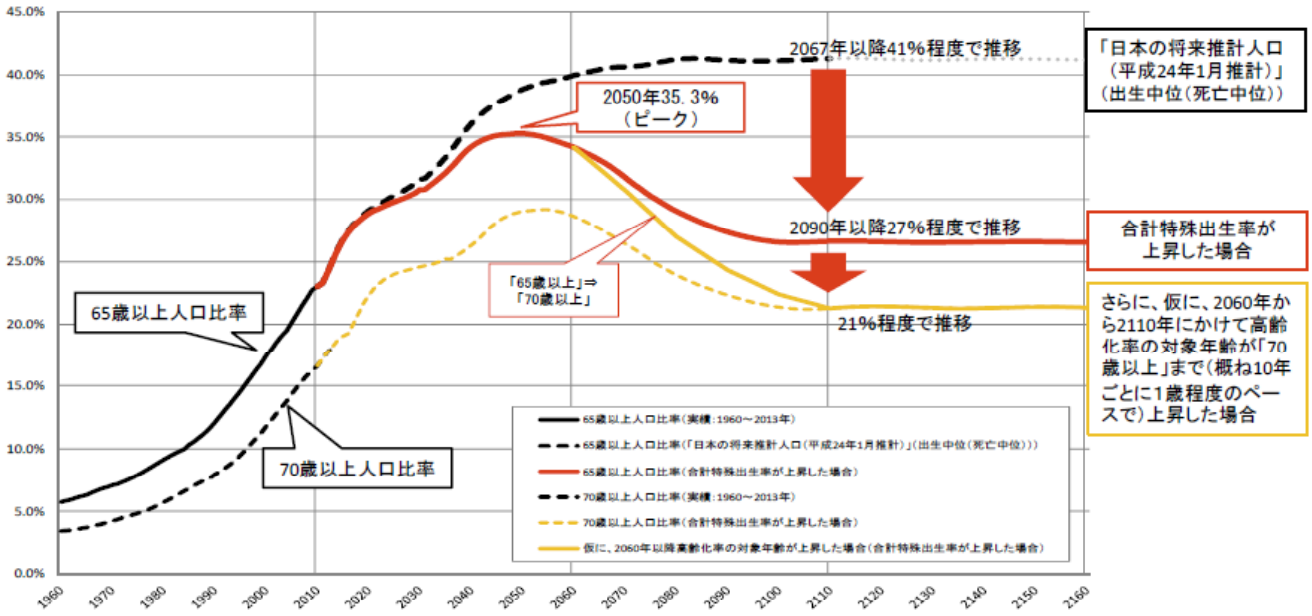
（注１）実績は、総務省統計局「国勢調査」等による（各年10月１日現在の人口）。国立社会保障・人口問題研究所「日本の将来推計人口（平成24年1月推計）」は出生中位（死亡中位）の仮定による。2110～2160年の点線は2110年までの仮定等をもとに、まち・ひと・しごと創生本部事務局において機械的に延長したものである。
（注２）「合計特殊出生率が上昇した場合」は、経済財政諮問会議専門調査会「選択する未来」委員会における人口の将来推計を参考にしながら、合計特殊出生率が2030年に1.8程度、2040年に2.07程度（2020年には1.6程度）となった場合について、まち・ひと・しごと創生本部事務局において推計を行ったものである。

図２．我が国の高齢化率の推移と長期的な見通し

創生

○　「日本の将来推計人口（平成24年1月推計）」（出生中位（死亡中位））では、高齢化率（65歳以上人口比率）は、将来的に41%程度まで上昇すると見通されているが、仮に、出生率が上昇すれば、2050年の35.3%をピークに、長期的には、27%程度まで低下するものと推計される。

○　さらに、将来的に健康寿命の延伸等に伴って高齢化率の対象年齢が「70歳以上」まで上昇するとすれば、高齢化率（70歳以上人口比率）は、概ね21%程度まで低下することとなる。



（注1）実績は、総務省統計局「国勢調査結果」「人口推計」による。国立社会保障・人口問題研究所「日本の将来推計人口（平成24年1月推計）」は出生中位（死亡中位）の仮定による。2110～2160年の点線は2110年までの仮定等をもとに、まち・ひと・しごと創生本部事務局において機械的に延長したものである。
（注2）「合計特殊出生率が上昇した場合」は、経済財政諮問会議専門調査会「選択する未来」委員会における人口の将来推計を参考にしながら、合計特殊出生率が2030年に1.8程度、2040年に2.07程度（2020年には1.6程度）となった場合について、まち・ひと・しごと創生本部事務局において推計を行ったものである。

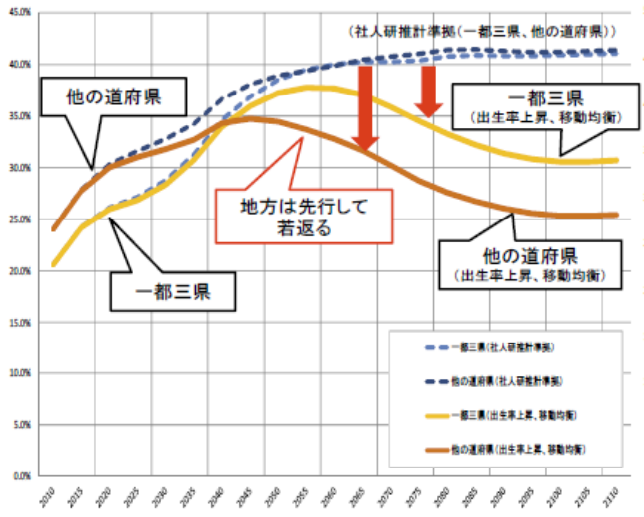
図３．地域別の高齢化率の長期的な見通し

創生

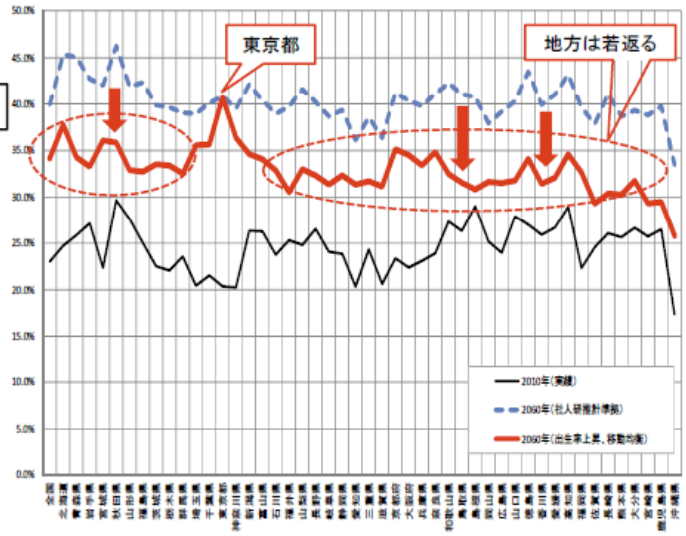
○　現状のまま推移したとすれば、一都三県においても、他の道府県においても、2070～80年頃以降、高齢化率は41%程度で推移するものと推計される。

○　仮に、2040年までに、現行程度の地域間の出生率格差を残しつつ全国の合計特殊出生率が2.07程度まで上昇し、移動が均衡した場合には、高齢化率は、他の道府県では、2045年頃の35%程度をピークに25～26%程度まで低下、一都三県では、2055年頃の38%程度をピークに30～31%程度まで低下すると推計される。

一都三県、他の道府県別　高齢化率（65歳以上人口比率）の見通し



都道府県別　高齢化率（65歳以上人口比率）の見通し（2010年→2060年）



（注1）2010年（実績）は、総務省統計局「国勢調査結果」による。
（注2）「社人研推計準拠」は、国立社会保障・人口問題研究所「日本の地域別将来推計人口（平成25年3月推計）」の2040年までの傾向を延長して、まち・ひと・しごと創生本部事務局において推計したもの。性・年齢階級別人口が同研究所の「日本の将来推計人口（平成24年1月推計）」（出生中位（死亡中位））の値に一致するよう補正を行っている。
（注3）「出生率上昇、移動均衡」は、上記「日本の地域別将来推計人口」のデータを用いて、現行程度の地域間の出生率格差を残しつつ、全国の合計特殊出生率の水準が2030年に1.8程度、2040年に2.07程度と上昇し、かつ、2040年までに移動が均衡した場合（純移動率がゼロとなった場合）について、まち・ひと・しごと創生本部事務局において推計を行ったものである（全国の推計値で補正を行っている）。

「日本創成会議」（日本生産性本部）の提言（平成26年5月）

創生

○ 日本の希望出生率（1.8）を2025年に達成することを基本目標。
○ 20歳代の結婚割合増加が、目標達成のための鍵。

基本目標＝国民の『希望出生率』の実現

◆現状（2013年）*出生率＝1．43*

◆基本目標（2025年）

『希望出生率』＝1．8

●国民の「希望出生率」として出生率＝1．8を想定。
・夫婦の意向や独身者の結婚希望等から算出。
・最も出生率が高い沖縄県は出生率＝1.8～1.9
・OECD諸国の半数が出生率＝1.8を超えている。

（参考）人口置換基準 *出生率＝2．1*

●将来人口が安定する「人口置換水準」は2．1
・日本の夫婦の理想平均子ども数は2.42人
・米、仏、英、スウェーデンの出生率は2前後

＜出生率向上の要因＞

1．結婚割合の上昇

◎20歳代～30歳代前半に結婚・出産・子育てしやすい環境を作る

○出生率1．8
・20歳代後半の結婚割合（現在40％）が60％になれば実現可能

○出生率2．1
・20歳代前半の結婚割合（現在8％）が25％に、20歳代後半が60％になれば実現可能。

2．夫婦の出生数増加

◎第2子、第3子以上の出産・子育てがしやすい環境を作る

具体的施策

若年世代の経済基盤確保
○雇用・生活の安定化

結婚・妊娠・出産の支援
○出会いと結婚の機会づくり

子育て支援
○待機児童の解消

働き方改革
○育児休業の拡充

多子世帯への支援
○税・社会保障制度見直し

政策総点検

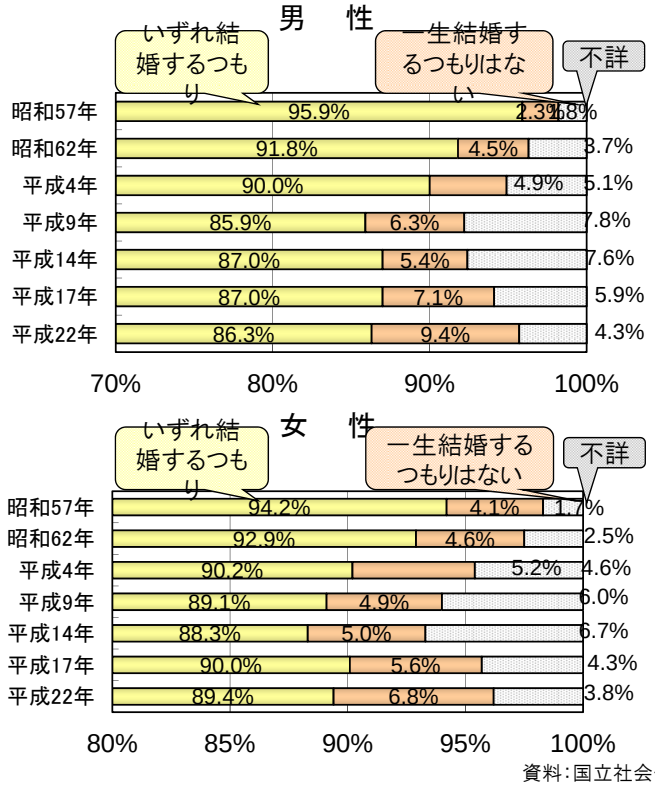
高齢者政策の見直し

国民の結婚や出産に対する希望

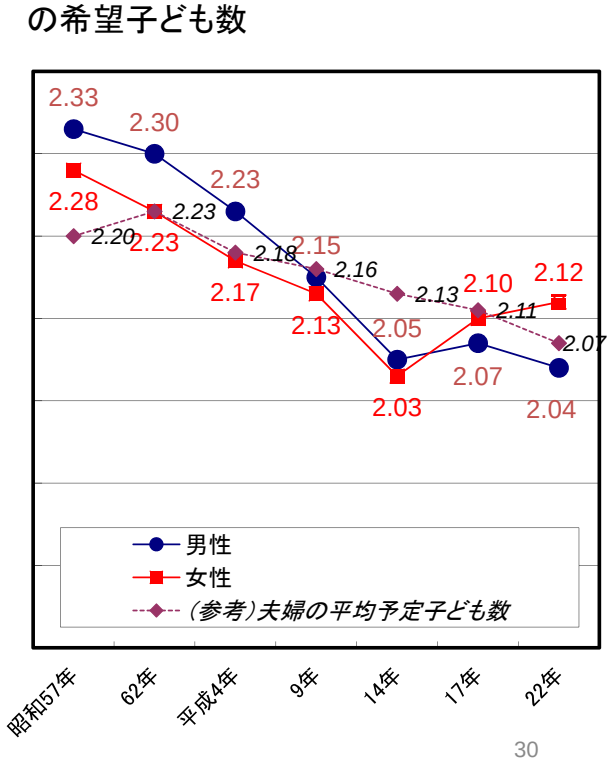
創生

独身男女の約9割は結婚意思を持っており、希望子ども数も男女とも2人以上。

○「生涯の結婚意思」について



○「いずれ結婚するつもり」の未婚男女の希望子ども数

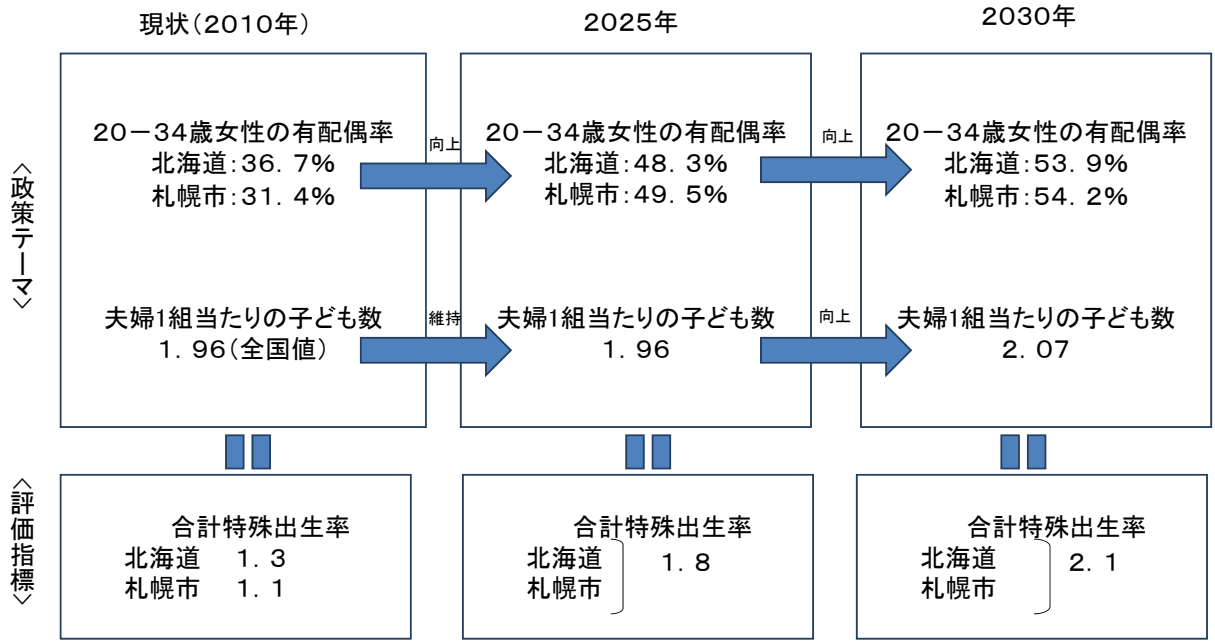


資料：国立社会保障・人口問題研究所「出生動向基本調査」（調査対象は18～34歳の未婚者）

「地域人口ビジョン」の作成例（人口数値目標：北海道、札幌市）

○ 20－34歳女性の有配偶率が10%ポイント～20%ポイント改善することで、出生率は希望出生率（1.8）に到達する見込み。
○ さらに夫婦1組当たりの子ども数が上昇すれば、人口置換水準の出生率（2.1）まで出生率が改善することも可能。

人口数値目標（例）



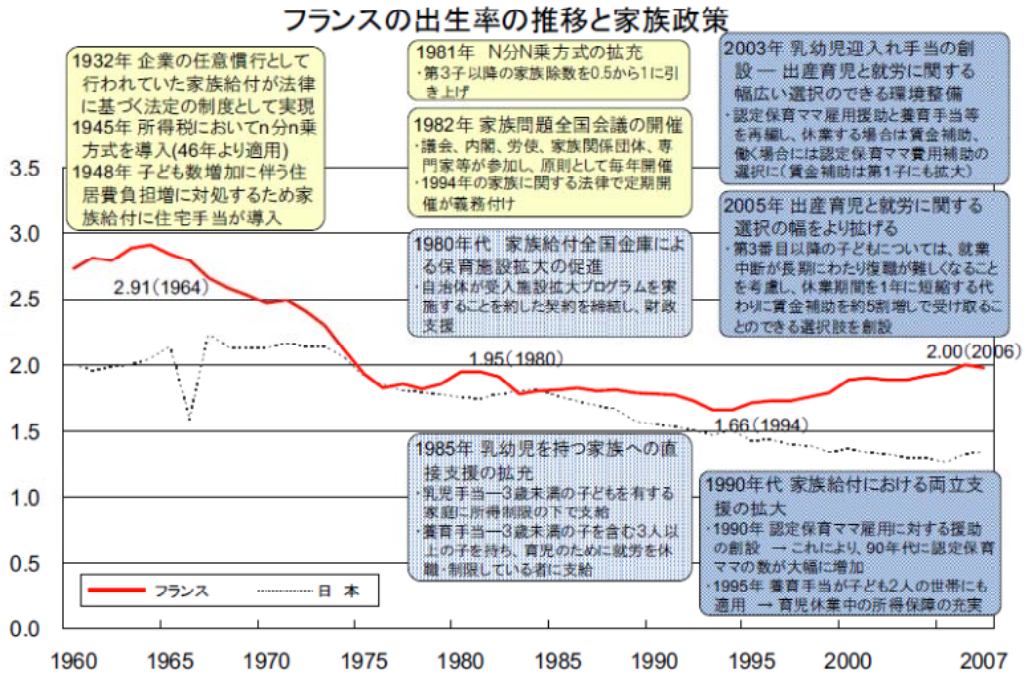
少子化への対処

政策的対応

フランスの少子化対策

創生

○ フランスでは、80年以上以前から少子化対策を実施しており、1980年代以降徐々に出生率が上昇。

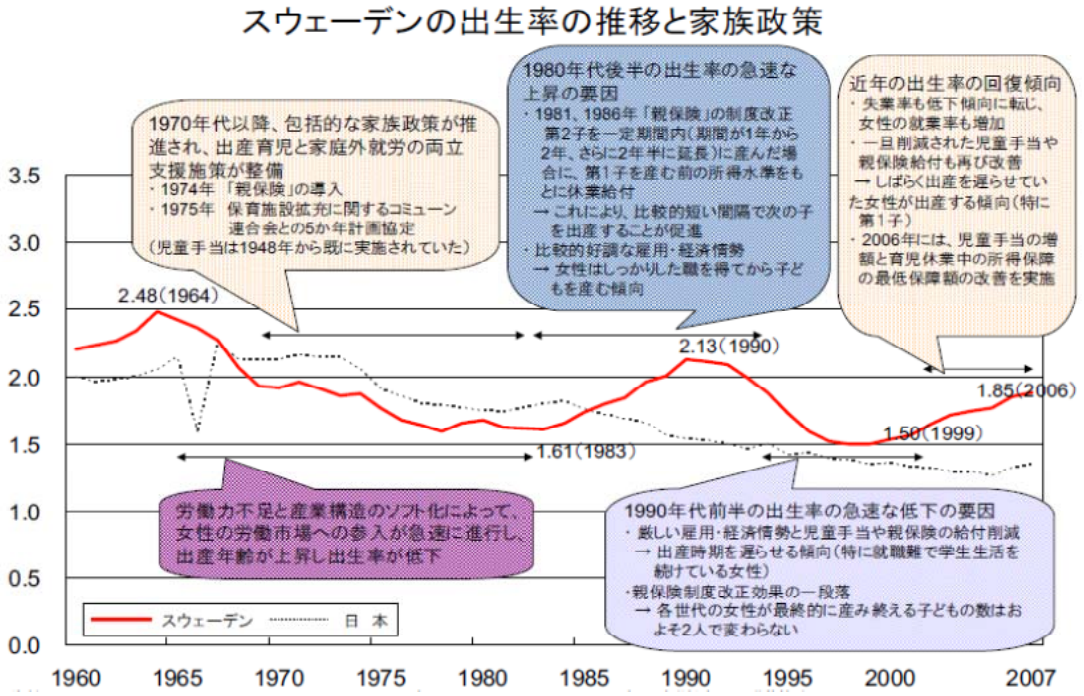


(出所)明治大学・加藤久和教授資料

スウェーデンの少子化対策

創生

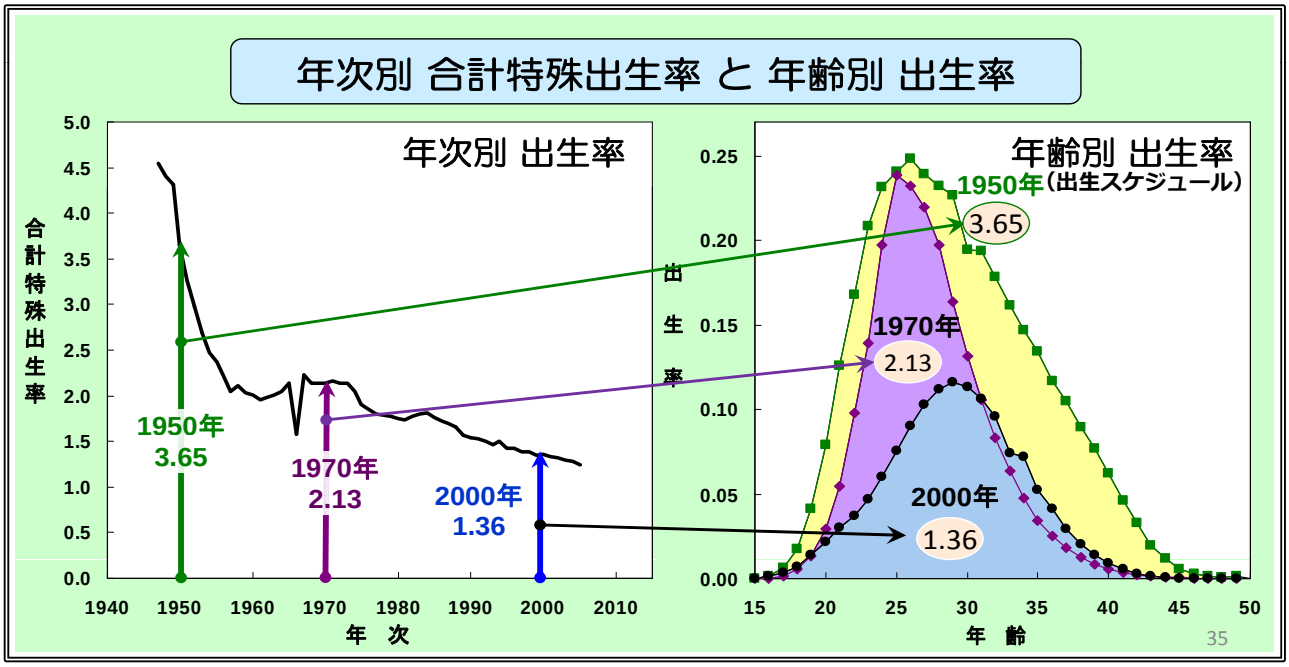
○ スウェーデンは、直近50年で二度の出生率低下を経験するも、休業給付や児童手当をはじめとする政策対応により改善。



合計特殊出生率と年齢別出生率（出生スケジュール）

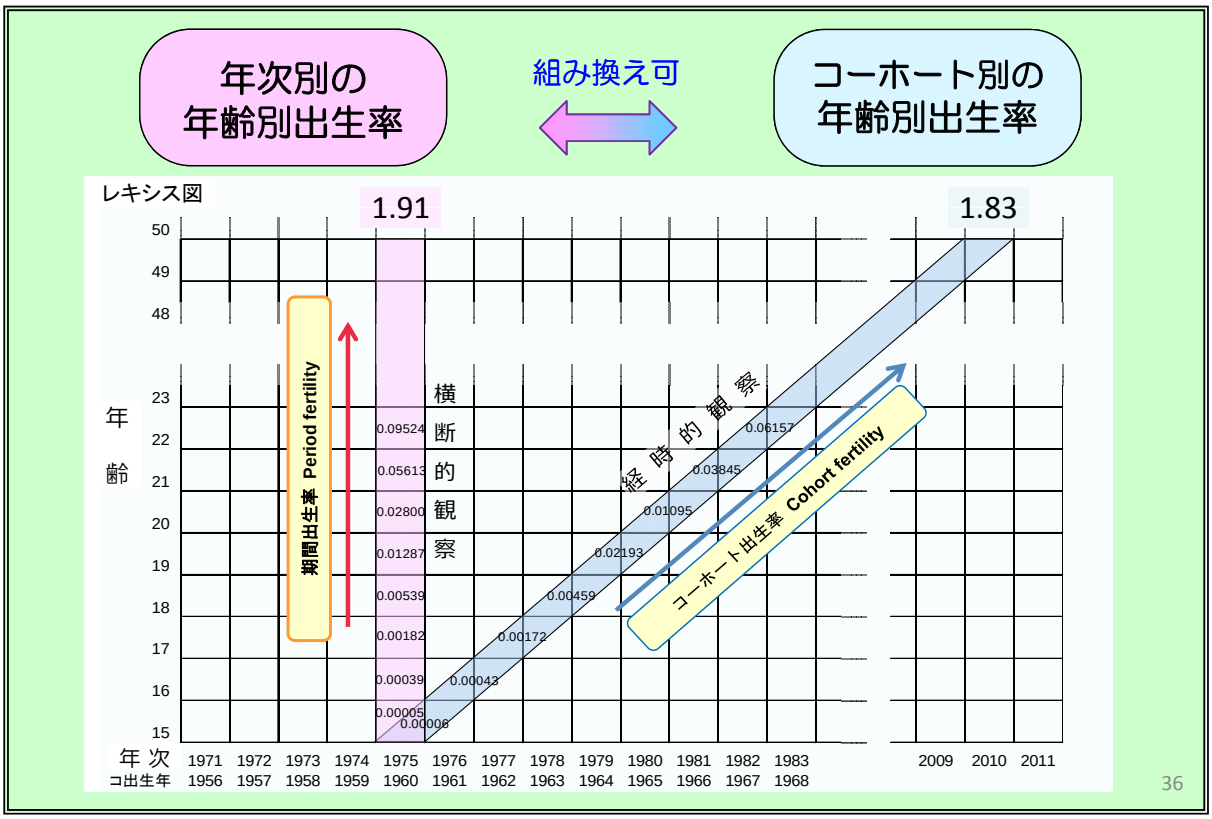
合計特殊出生率とは、
- ある年次に観察された女性の **年齢別出生率** を、全年齢にわたって合計した数値.
- それは，女性がその年齢別出生率にしたがって子どもを生んだ場合の 生涯の出生子ども数 と解釈できる.
- 年齢別出生率は、出生スケジュール とも呼ばれ、(女性の)ライフコースを反映する。

下図右： 1950年頃までの日本女性の50歳までの人生は、妊娠・出産・子育てに終始
→ 1970年頃までに30歳代、10代後半は妊娠・出産から解放　→ その後、20歳代も妊娠・出産離れ ＝ 少子化



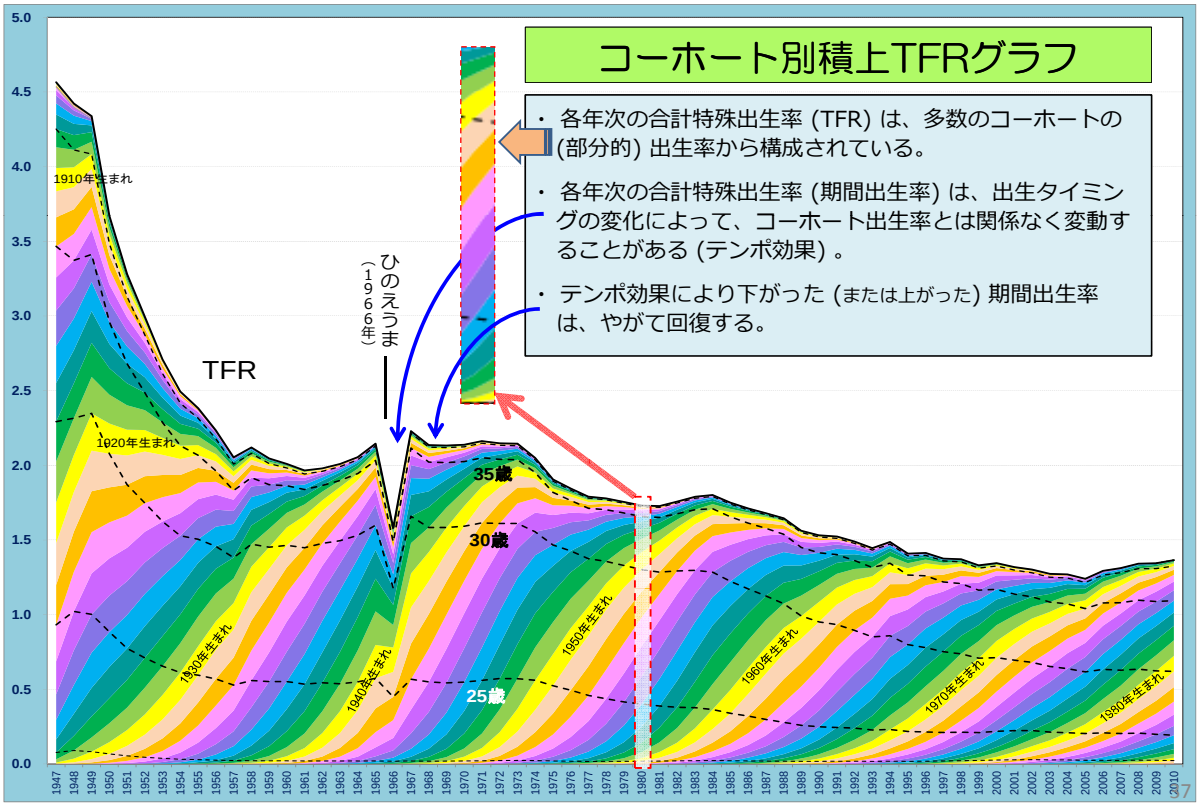
２種類の出生スケジュールと合計特殊出生率

- 出生スケジュールおよび合計特殊出生率の観察軸には２種類ある。→ **年次(期間)** 合計特殊出生率 vs **コーホート** 合計特殊出生率

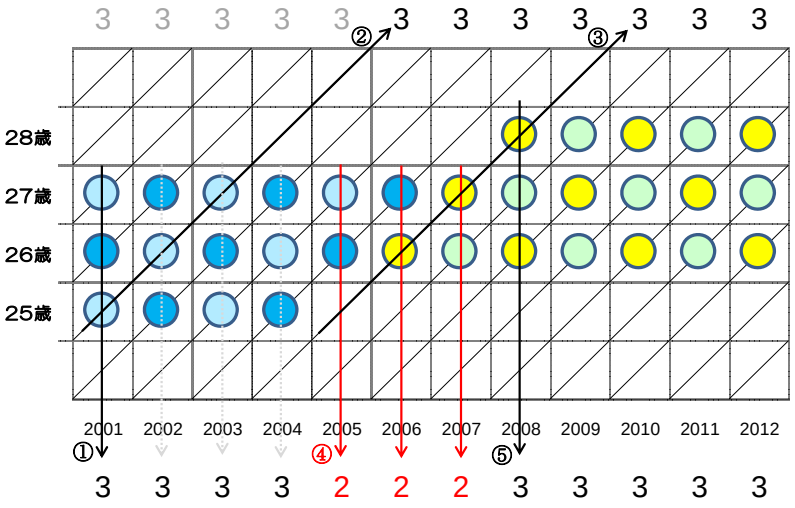


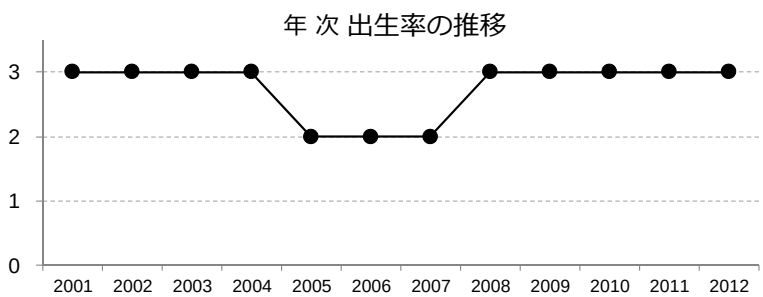
年次(期間)出生率の成り立ち

・各年次の合計特殊出生率は、必ずしもライフコース的整合性のない多数のコーホートの出生率から構成されている。



“晩産化”の 年次(期間)出生率に対する効果





“晩産化”の実験

① 年 次 2001年の出生数は 3 人

② コーホート[2001年25歳]の生涯の出生数は 3 人

※この例ではすべてのコーホートの生涯出生数は 3 人

③ コーホート[2005年25歳]が 1 年、晩産化（しかし生涯の出生数は 3 人）

④すると 年 次 2005年の出生数は 2人に減少！

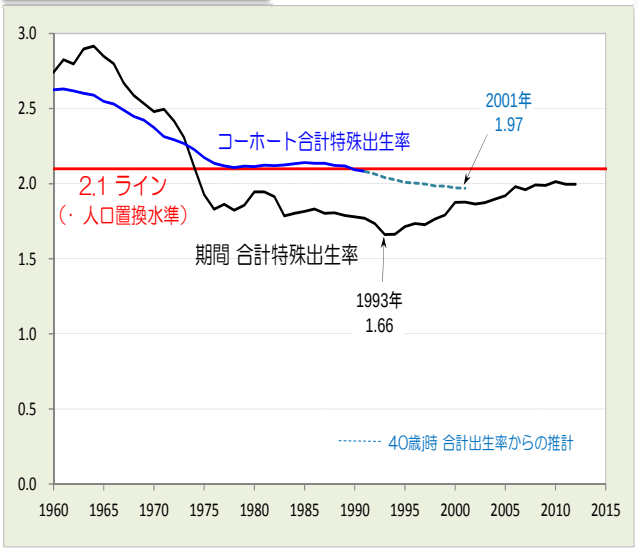
⑤ コーホートの晩産化が止まるとピリオドの出生数が 3 人に回復。

コーホート出生率が安定していても、「晩産化」が起こると年次の出生率は低下する。

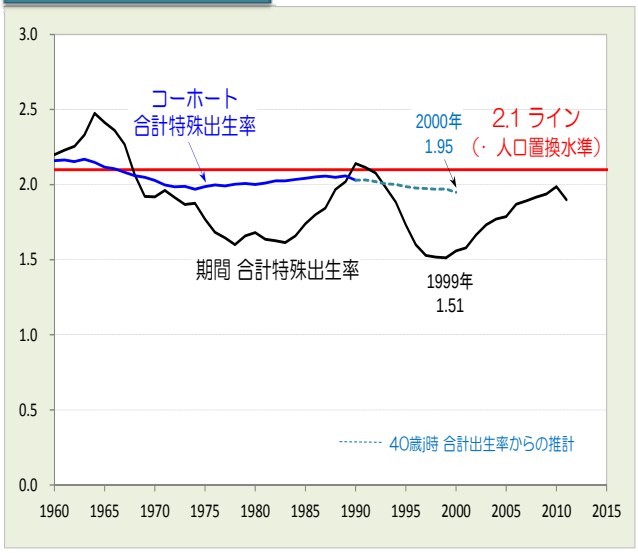
年次的に観察される出生率の注意点

フランスやスウェーデンは、施策の充実等によって少子化の状態（人口置換水準下の出生率）から脱し、出生率を回復させた国の例として取り上げられることがある。しかし、これらの国でみられた年次的な出生率（期間合計特殊出生率）の低下は、主に出生年齢の遅延にともなう効果（テンポ効果）によるもので、その背景にある実質的な出生力（コーホート合計特殊出生率 ≅ 生涯の平均出生子ども数）はもともと人口置換水準付近で推移していた。したがって、近年の出生率回復はこのテンポ効果が弱まったことが原因であり、人々の持つ実質的な子ども数が増加したわけではない。

フランス



スウェーデン



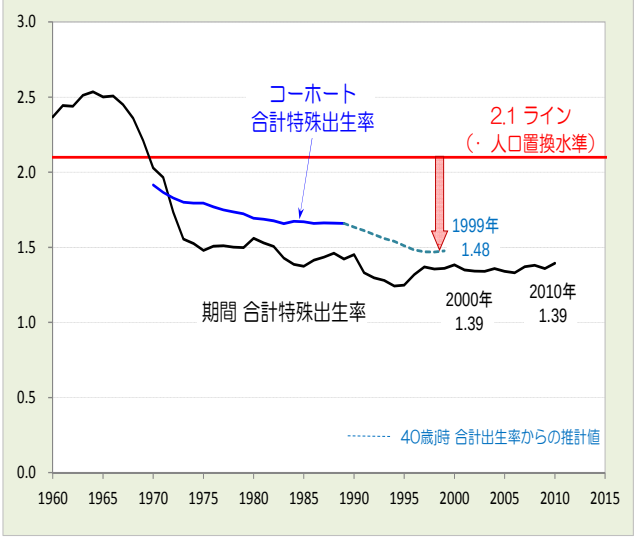
資料：The Human Fertility Database (http://www.humanfertility.org), 取得年月日：2014/12/17 より作成.

年次的に観察される出生率の注意点

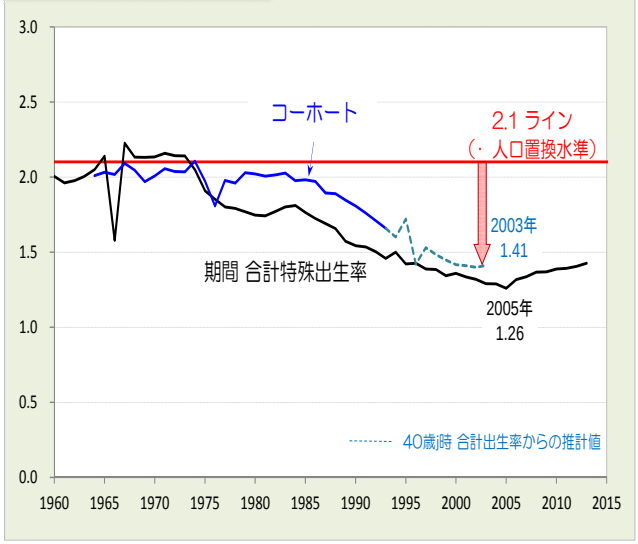
これに対して、ドイツ や 日本の出生率低下（下図）は、かなりの部分が実質的な出生力（コーホート合計特殊出生率）の低下に起因しており、今後、少子化が解消するためには（出生率が人口置換水準へ復帰するためには）、フランスやスウェーデンとは異なる特段の努力が必要となる。

★ 少子化を正しく捉えるには、指標の背後にある ”実質” を読み取る必要がある。
→ 少子化に正しく対処するには、指標の示す出生の “量” ではなく、”質” に着目する必要がある。

ドイツ



日　本

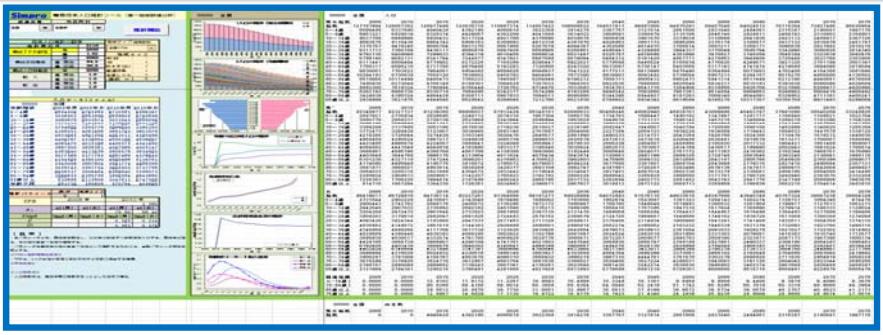


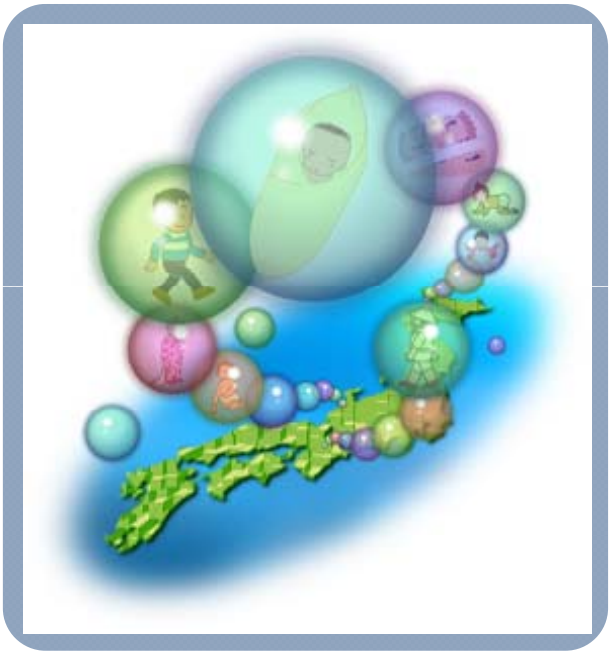
資料：The Human Fertility Database (http://www.humanfertility.org), 取得年月日：2014/12/17 より作成.

地方創生：統計データツールの開発・利用

地域経済分析システム：RESAS

小地域将来人口・世帯数推計システム：Simpro (仮)





地方創生と人口統計

－ まち・ひと・しごと再生の課題 －

科研「経済統計・政府統計の理論と応用」

2016年 1月29日（金）

東京大学経済学部小島ホール

国立社会保障・人口問題研究所
金子隆一

地域少子化対策検証プロジェクト（第２回会合 2015 年 10 月 22 日(金)）
提出資料５補足説明

少子化指標としての出生率について

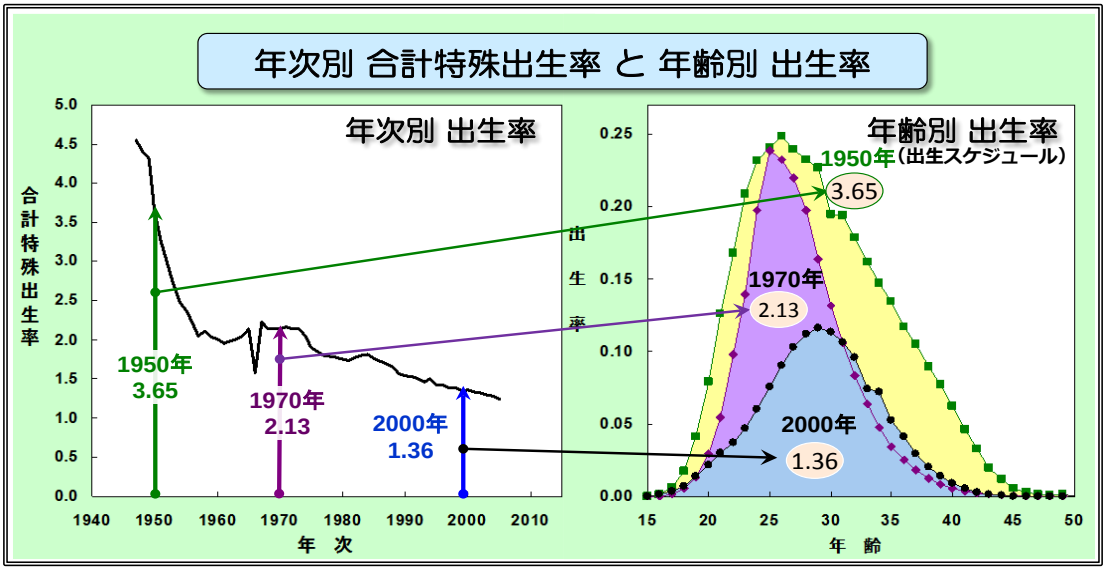
金子隆一（国立社会保障・人口問題研究所）

少子化対策を考える際には、統計指標によって少子化の現状を正しく把握することが必要となるが、その際に指標の性質に関する十分な理解を欠くと、誤った現状認識に陥ることがある。出生率（合計特殊出生率）は、少子化を論ずる上で最も基本的な指標と考えられるが、実はその見方についてはいくつか注意すべき点が存在する。すでに広く使われている指標ではあるが、ここでは一旦基礎に立ち戻って、この出生率の成り立ちや性質について解説し、その正しい見方や少子化対策における使い方について考えて見たい[1]。

１．出生率と出生スケジュール

図1の左には、一般によく見られる合計特殊出生率の年次推移を示した。この年々の出生率の元となるのは、年齢別出生率であり、これをグラフとして描いたものを同図右に示した。横軸は女性の年齢、縦軸は各年齢における出生の頻度（すなわち年齢別出生率）であり、女性の再生産年齢とされる 15 歳から 50 歳までの変化を描いている。これは「出生スケジュール」とも呼ばれる。この出生スケジュールを観察することで、女性が人生の中でどのように出産しているかがある程度見えてくる。

図１　合計特殊出生率と年齢別出生率（出生スケジュール）



資料：厚生労働省「人口動態統計」による.

[1] 本稿は、人口統計の専門家でない一般の読者を想定したものであり、一部の説明において専門的厳密性よりも、わかりやすさを優先した表現を用いている。

同図には、３つの時期のスケジュールを示しているが、一番後ろに見える黄色のグラフは 1950 年の女性の出生スケジュールである。それは 10 代の終わりに急速に立ち上がり、40 歳代前半に至るまで高い水準を示しており、当時広い年齢範囲で出生が有ったことを示している。全体のレベルを表す合計特殊出生率は 3.65 と、戦前の水準（4～5）よりは低いものの、この頃までの日本女性では、成長直後の 10 代終わりから 50 歳頃までの人生はほとんど妊娠・出産・子育てと共にあったことが読み取れる。

これに対して、1970 年のグラフ（紫色）では、10 代ならびに 30 代以降の出生がすっかり消失しており、女性がこれらの年代において、妊娠・出産から解放されたことがわかる。これは当時、女性に高学歴化と社会進出が始まったことを反映しており、またその後の進展の前提ともなっている。なお、出生スケジュールの曲線が囲む面積は、合計特殊出生率に相当するため、このような年齢別出生率のグラフによって、各時期の出生水準を視覚的に比較することができる。左図に見た出生率の年次推移は、この面積の時系列変化を示したものに他ならない。

続けて 2000 年の出生スケジュールのグラフを見ると、20 代で劇的に「出産離れ」が起きており、これによって全体の面積が大幅に削り取られ、少子化が進んだ姿を顕している。このように、出生スケジュールを見ることで、単一の数値である合計特殊出生率の背後にある女性のライフコースの姿を垣間見ることができる。

２．２つの合計特殊出生率：年次観察とコーホート観察

人口学では、官庁統計として年次ごとに把握されるマクロの人口統計（国勢調査結果や人口動態統計報告など）から、人々のライフコースを再構築して分析することが行われるが、その際に用いられるツールの一つとして「レキシス図」と呼ばれるものがある。それは**図2**に示したように、横軸に年次、縦軸に年齢を配置し、格子を描いた座標平面である。

この平面上で、たとえば 1960 年に生まれて 75 年に 15 歳を迎えた世代のライフコースは、ブルーで示した斜めの帯として描かれる。右方向に１年進めば、個人の年齢は上方に１年上がることになるので、彼らの人生は右斜め 45 度上方に向けて進む。

図では、その途中に年齢別出生率を数値として示してあるが、これを斜めに足し合わせた値 1.83 がこのコーホートの合計特殊出生率である[2]。これはこのコーホートの生涯の平均出生子ども数に相当し、かれらの出生の量的水準を表す指標となる[3]。したがって、世代ごとに少子化がどのように進んでいるかなどを測るには適した指標となる。

しかしながら、この指標は生涯の出生を終える年齢（便宜的に 50 歳と定義）に達した世代についてしか計算できないため、現在の状況を知るためには少なくとも出生の主力となっている 25～35 歳世代が 50 歳に達する 25 年後まで待たなければならない。

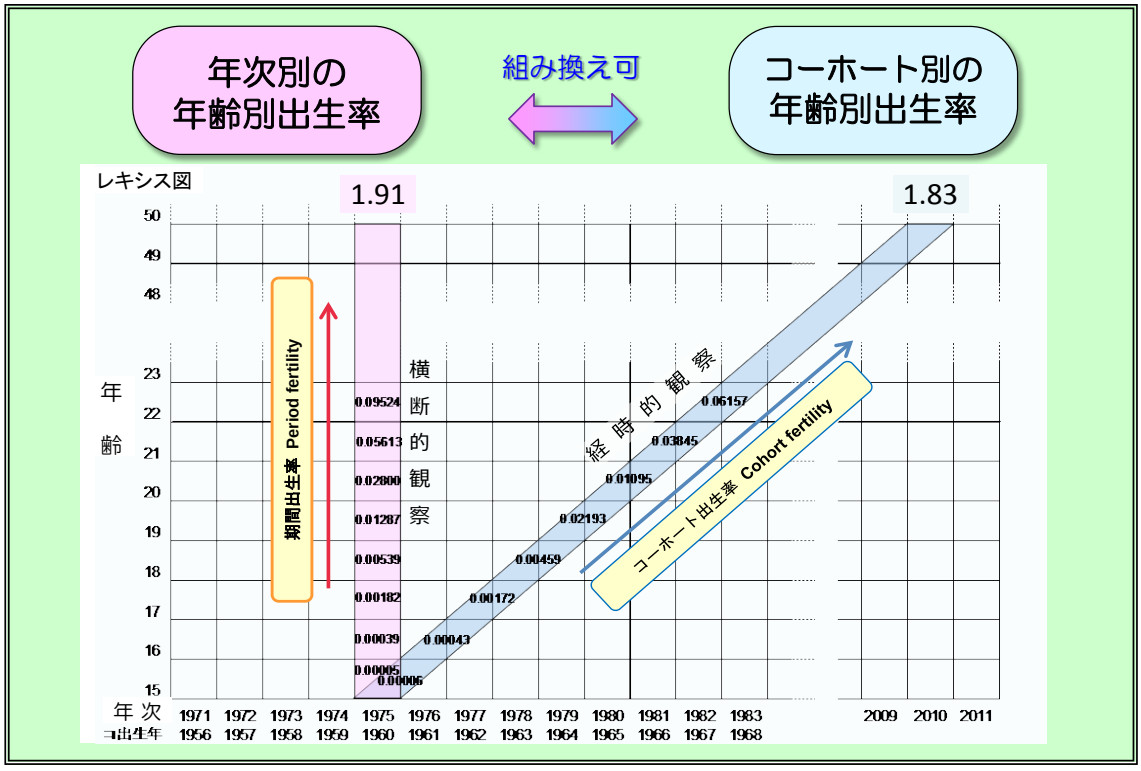
---

[2] コーホートとは、ここでは同じ年に生まれた人々の集合を指す。「世代」とほぼ同義。

[3] 厳密には、合計特殊出生率は、「再生産期間を通して女性集団に死亡が発生せず、所与の年齢別出生率に従って子どもを生んだ場合の生涯の平均出生子ども数」を与える。

これに対して、たとえば去年とか今年などのように、ある１年次の出生率を知るためには、**図2**においてピンクの帯で示されるような横断的観察を行うことが考えられる。15 歳から縦上方向に年齢別出生率を足して行けば、この年次の出生スケジュールにしたがって子どもを生んだ場合に期待される生涯の平均子ども数の値が算出できる。実は、これが毎年報道され、一般に用いられている「合計特殊出生率」であり、図の例では、「1975 年の合計特殊出生率は、1.91 である」と表現されることとなる。

図2　2種類の出生スケジュールと合計特殊出生率



資料：数値（出生率）は厚生労働省「人口動態統計」による.

このように合計特殊出生率には、世代（コーホート）のライフコースに沿って算出される「コーホート合計特殊出生率」と、対象年次の年齢別出生率を足し上げた「年次の合計特殊出生率」の２種類が存在する[4]。前者は実在する世代の生涯（ブルーの帯）について算出した平均出生子ども数であるのに対して、後者は多数の世代の部分的出生率を合成して算出した、いわば「仮想的なライフコース」（ピンクの帯）の平均出生子ども数となる。

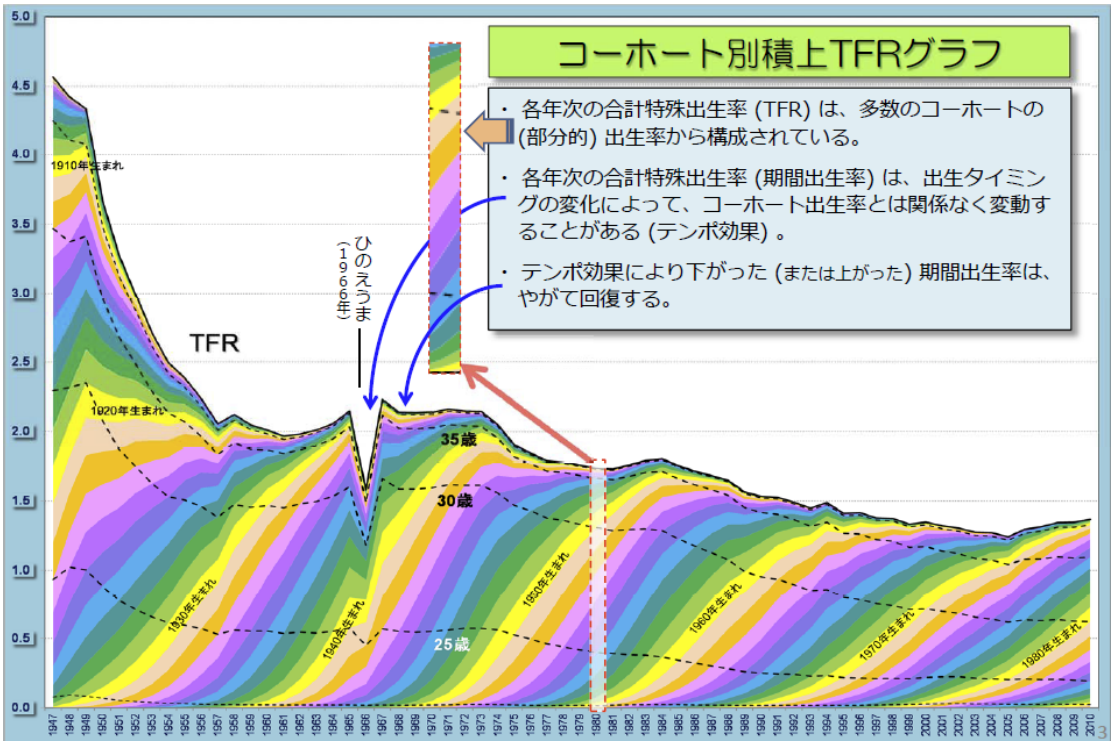
３．合計特殊出生率の成り立ちと性質

図3には、我々がよく目にする「年次の合計特殊出生率」の推移を、その内訳となるコーホートの年齢別出生率（出生スケジュール）で色分けして描いた。ある一年次を抜

[4] ここで年次の合計特殊出生率と呼んでいる指標は、専門的には「期間合計特殊出生率」あるいは「ピリオド合計特殊出生率」と呼ばれている。

き出すと、それは多数のコーホートの部分的出生率によって構成されていることがわかる。

図3　2種類の出生スケジュールと合計特殊出生率



資料：厚生労働省「人口動態統計」による.

ここで問題となるのは、この抜き出した年次の年齢別出生率のセットは、一応すべての年齢の出生情報を含んでいるが、各年齢での出生行動は、実際のライフコースのように整合的であるとは限らないということである。たとえば、「晩産化」（本稿では、生涯に生む子ども数を変えずに出産年齢が遅くれることを意味するものとする。以下同じ。）が生じたとき、実際のライフコースでは 20 代など若年での出生率は低下し、その分を 30 代以降の高年齢で出生率が上昇することで取り戻すので、コーホート合計特殊出生率は変わらない[5]。ところが、年次の合計特殊出生率では、各年齢の出生率の間にそのように協調して動くしくみはないため、その時々の情勢に反応して、全年齢で一斉に上昇したり、低下したりすることの方が多くなる。

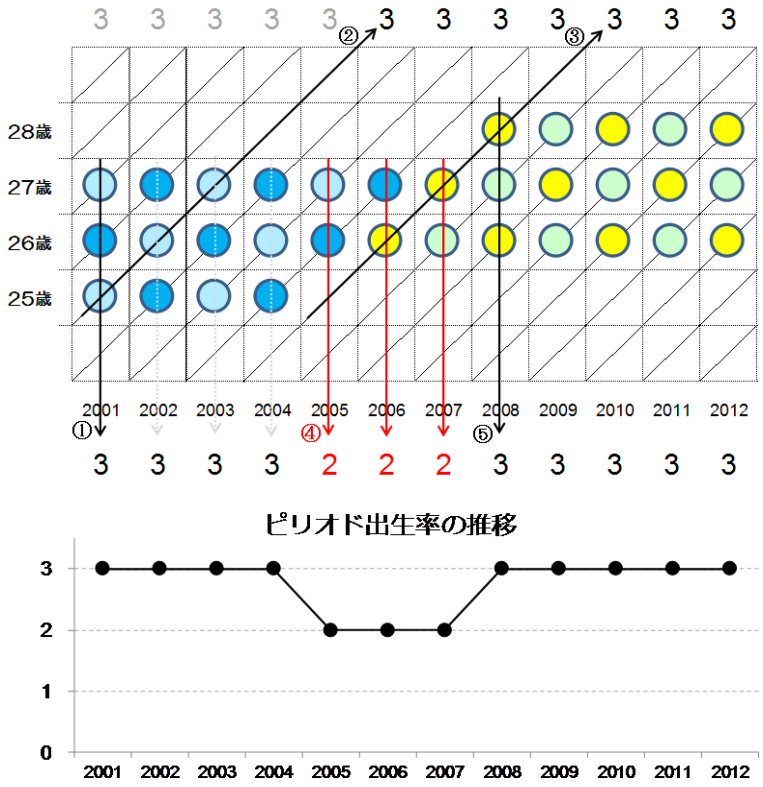
「ひのえうま」（1966 年）を例にすると分かり易い。この年に生まれた女児は不幸になるとの迷信から、人々がこの年の出生を翌年以降に先延ばしした結果、年間の合計特殊出生率は 1.58 と例年の３／４程度にまで低下した。しかし、この現象によって日本人女性が生涯にもつ平均の子ども数が３／４に縮小したわけではない。実際は、「ひのえうま」の年に子どもを生んでいたすべての世代のコーホート合計特殊出生率（平均

[5] 現実には出産年齢が高まると妊孕力（にんようりょく）が低下するため、高年齢での出生の取り戻しは完全ではなく、コーホート合計特殊出生率は幾分減ることが一般的であるが、ここでは簡単のため出産年齢の遅れだけのケースで説明する。

出生子ども数） は、前後の世代同様にほぼ２人となっていた。

この例は、人々が実際に持つ子ども数は変わらないのに、年次の合計特殊出生率は、見かけ上低下し得ることを示しており、この場合には「一人の女性が生涯に生む平均の子ども数」という合計特殊出生率の解釈が成り立たないことを意味する。「ひのえうま」の例では、このことはむしろ明白であろうが、日本の少子化過程で見られるような長期にわたる晩産化においても、程度の差こそ違え、同様のメカニズムが働いていたことはあまり知られていない。分かり易く表現するならば、日本の少子化過程（合計特殊出生率低下過程）においては「ミニひのえうま」が毎年生じており、人々が実際に生涯に生む子ども数と年次合計特殊出生率の値との間には齟齬が生じていたのである[6]。

図４ “晩産化”による年次の合計特殊出生率低下の実験



このことを簡単な実験により説明しよう。**図4**には、女性が 25 歳から 27 歳まで毎年一人ずつ出生するような仮想の集団を考え、ある世代以降、出生スケジュールが１年「晩産化」する例を示した。この例ではスケジュールが遅れるだけなので、すべての世代の最終的な子ども数は３人である（図上部の数値）。しかし、晩産化が右斜め上方への移動であることから、縦方向の合計値には一時的に「ずれ」が生ずる。この結果、晩産化過程の３年間については、合計特殊出生率が２人に低下することが示されている（図下部の数値とグラフ）。そして、この過渡期の直後、世代の出生スケジュールには

[6] 後述するように、日本の少子化においては「ミニひのえうま」効果とともに、実際の出生子ども数の低下が同時に進行しており、年次の合計特殊出生率の低下はこれらが複合したものであった。

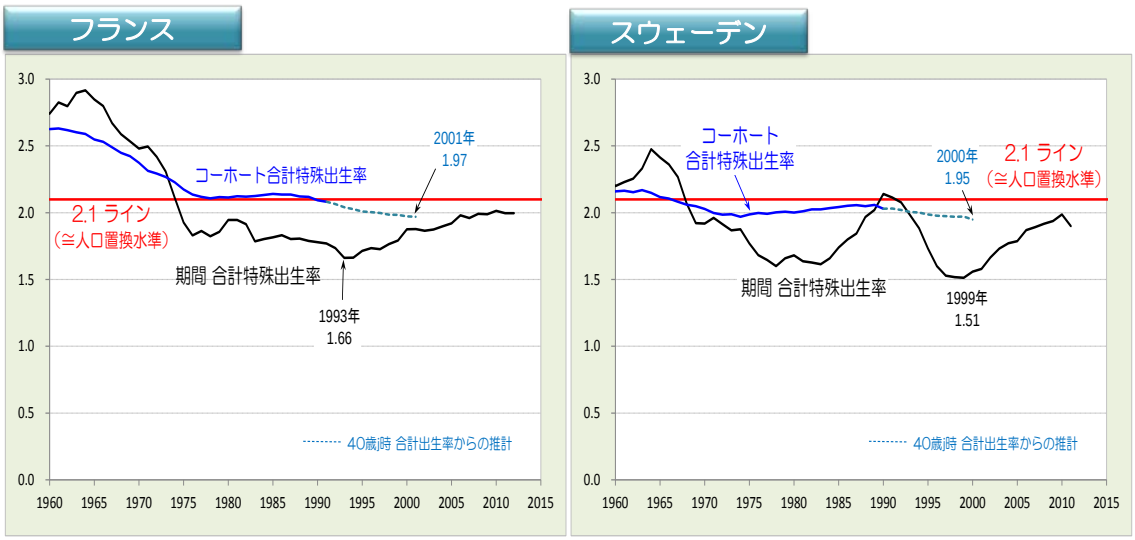
もはや変化がないにも関わらず、年次合計特殊出生率は自動的に３人に回復している[7]。

さて、以上の説明で「コーホート出生率が安定していても、晩産化が起こると、年次の出生率は低下する」という性質が確認できた。とすると、年次の合計特殊出生率は、その額面通りの数値を、少子化の動きとして理解することが適当でない場合が有ることになる。以下にその様な実例を取り上げて、問題点を検討してみよう。

４．少子化指標としての合計特殊出生率の問題点

図5には、フランスとスウェーデンについて、それぞれ２種の合計特殊出生率の推移を描いた。年次の合計特殊出生率の変化は黒の曲線で描かれている。これらを見ると、いずれの国でも 1970 年代後半以降において年次出生率が大きく低下し、人口置換水準（赤の線）から離れた時期が有ることがわかる。スウェーデンにおいては、90 年前後の回復を挟んで、２回の大きな低下を経験した（この著しい変動は「ジェットコースター出生率」と呼ばれたことがある）。

図5　フランス、スウェーデンの年次出生率とコーホート出生率



注：コーホート合計特殊出生率（青の曲線）は、各コーホートが 29 歳であった年次にプロットしている。
資料：The Human Fertility Database (http://www.humanfertility.org), 2014/12/17 データ取得.

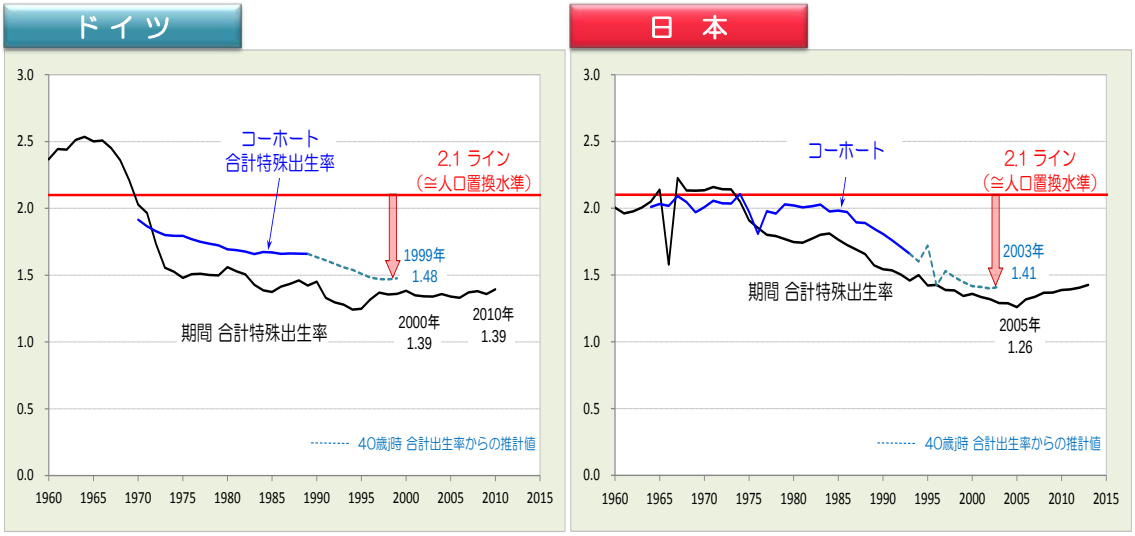
ただ、どちらの国においても 90 年代後半（あるいは末）以降は、年次出生率が回復基調にあり、人口置換水準付近にまで上昇していることが見られる。これらの出生率の回復は、低出生に喘ぐ他の国々からみると目覚ましいものであり、また、両国とも手厚い家族政策が実施されていることで知られているため、しばしば政策効果によって「少子化を克服」した例として紹介をされている。

しかしながら、同図に青色で示したコーホート出生率の推移を見ると、実はこの ２ヶ国とも、ずっと人口置換水準付近で推移しており、これは年次出生率に大きな低下が

[7] ちなみに年次出生率が低下した期間に失われた出生は、晩産化が終了して出生率が回復しても、もはや戻っては来ない。不思議に思うかもしれないが、それは晩産化による世代間隔の延長の代償として消費された形となる。

見られた時期においても、実質的な出生力は低下していなかったことを示している。すなわち、年次出生率の一時期の低下は、先に紹介した晩産化による見かけ上の現象（ミニひのえうま効果）であったことになる[8]。つまり現時点で振り返ってみると、当該時期の出生率の回復は、この晩産化が緩んだことによるミニひのえうま効果の解消が主な理由であって、必ずしも人々の持つ子ども数が増えたことを示すものではない。要するに、両国のこの時期の年次出生率の回復は、私たちが単純に考える「少子化を克服」したという図式、すなわち一時期人々が生涯に持つ子どもの数を減らしたが、これに即応した施策等によってその後の世代では子ども数が復活してきたという物語とは、どうも違うようなのである[9]。だから、他の少子化国でこの両国の経験から学ぼうとするなら、最近の年次出生率を回復させたという事例ばかりに注目するのではなく、コーホート出生率を長期に人口置換水準に安定させている理由にこそ注目しなくてはならない[10]。

図6 ドイツ、日本の年次出生率とコーホート出生率



注：コーホート合計特殊出生率（青の曲線）は、各コーホートが 29 歳であった年次にプロットしている。
資料：The Human Fertility Database (http://www.humanfertility.org), 2014/12/17 データ取得.

---

[8] いわゆる「ミニひのえうま」効果は、専門的にはテンポ効果と呼ばれるものであり、厳密には晩産化などの平均出生年齢の上昇・下降だけでなく、分散変化などを含む出生年齢分布の種々の変化によって生ずるものであるが、ここでは簡単のため、両国の出生年齢分布に生じたこれらの変化をまとめて晩産化と呼んでいる。

[9] これらの例でも、出生率回復期において晩産化が緩んだ原因として政策効果を挙げることは可能である。政策の変更等はしばしば人々の出生年齢に大きな影響を与えることがあり、スウェーデンの「ジェットコースター出生率」は実はその典型的な例にあたる。ただし、それらが実質的出生力（コーホート合計特殊出生率）の変化をともなわないならば、単に出生の起こる年次が変わるだけであって、「少子化の克服」にはあたらない。それどころか、年々の出生数に上下動をもたらし、その後の人口の年齢構造に長期の不均衡をもたらすという点で、負の影響を残すことにすらなる。

[10] フランス、スウェーデンの出生率低下期について、後の晩産化の緩みがなければ、先送りされた出生がそのまま遺失されてしまった可能性があり、少子化国に転ずる危機にあったと考えれば、出生率回復期の政策的対応が少子化阻止に有意義であったとの見方ができるだろう。しかし、その場合であっても両国の長期にわたる手厚い家族政策の下支えが前提条件になっていたはずであり、短期的な政策のみに注目することでは、他の少子化国に応用可能な知見は得られないと考えられる。そのことは同時期の他の欧州諸国との比較などから導かれる。

これに対して、**図6**にドイツと日本に関する同様のグラフを描いた。この２国においても年次の合計特殊出生率（黒線）が 70 年代半ば頃から人口置換水準を離れ、大きく低下を示しているが、フランス、スウェーデンとの違いは、最近において低下こそ止まったものの、1.5 を下回る低水準に留まっており、置換水準に向けての回復の兆しが見られないことである。そして、より大きな違いは何といってもコーホート合計特殊出生率（青線）が、人口置換水準を大きく割り込んで低迷している点である。すなわち、この２ヶ国の年次出生率の動向は一時期のフランス、スウェーデンにおける年次出生率の低下とは実質的に異なるものであり、晩産化の緩みのみでは置換水準付近にまで回復することはできない。実際、日本においても 2006 年以降、女性 30 代後半以降の「駆け込み出産」という形の晩産化の緩みによって年次出生率の向上が見られているが、その水準は未だ置換水準には遠く及ばない[11]。

考察

さて、本資料では、合計特殊出生率の成り立ちについて解説するとともに、少子化の動向把握において年次別合計特殊出生率だけを見ていると、人々が生涯にもつ子ども数の動向に目が向かず、少子化の動向を正確に把握できない危険があることを紹介した。それでは、年次別の合計特殊出生率は、少子化を把握する指標として放棄すべきかといえば、決してそうではない。それは国や地域における特定の期間（年次）における人々の出生行動の強度について的確な数値を与えるものである。問題はそれが少子化の議論に必須な視点であるライフコースに沿った出生行動についての情報をもたらさないということと、もたらさないにも関わらず、「１人の女性が生涯に生む平均子ども数」という単純化したライフコース的解釈によって理解されている場合がある点にある。

つまり、年次別合計特殊出生率は、単年に起きた年齢別の出生行動を、理解しやすいように「生涯の平均出生子ども数」に相当する数値として提示してくれる親切な指標なのだが、晩産化など出生スケジュールに変化が起きていると、「ひのえうま」年の 1.58 のようにライフコース視点から見ると意味のない数値になってしまい、親切が仇となってしまうということである。しかも、現実の少子化は、晩婚化・晩産化などのタイミング変化と切っても切れないので、この指標による観察は常に誤解と背中合わせになる憾みがある。

残念ながらこうした問題点をすべてクリアする単一の指標はない[12]。そもそも「指標」というものは、本来複雑な現実の複雑性を捨象し、特定の部分だけを強調して表現するためのものであるから、その特性や限界をよく知り、目的に合わせて使い分ける必要がある。欠けている部分は、他の指標や情報を駆使することになる[13]。

---

11 筆者の考えでは仏、典の長期にわたる手厚い家族政策の継続的、安定的、発展的持続こそが、他の少子化諸国との違いを生んでいるものである。この点には、次節「考察」においても再び言及する。

12 年次の合計特殊出生率の晩産化にともなう変化をテンポ効果として分離し、これを差し引いた指標（Bongaarts-Feeney の調整合計特殊出生率など）も提案されているが、数値の解釈などに対して批判がある。

13 地方における少子化対策のために、まち・ひと・しごと創生本部から提供されている未婚率や出生順位

さて、以上では、少子化という現象を扱う上での出生率という指標の見方、使い方について人口統計学分野から解説と注意喚起を行った。以下では、そのような知見が、少子化対策に対して示唆することを、私見も交えつつ、もう一歩踏み込んで考えてみたい。すなわち、「出生率」という指標は、何を捨象し、何を強調しているのかということである。この指標が純化して提示しようとしているものは、ある時期の国や地域における出生の「量」である。平たくいえば、それはある年の出生の「頭数」の多寡を、「1人あたり」や「ライフコースあたり」の数値に換算して示しているものである。一方で、少子化対策が国民、とりわけ当事者となる「生まれてくる子ども達」と「その親世代」の福祉（well - being）のために行われるものであるなら、まず注視すべきものとしては、出生の「質」が中心となるべきである。それは倫理的にそうあるべきというだけでなく、少子化対策における個々の施策の実効性に対しても重要な意味を持つと考えられる。なぜならば、少子化対策を受ける側にとっては、それが「頭数合わせ」などではなく、その人々のライフコースの質を高めようとするものなのだという「メッセージ」が受信され、受容されたときにはじめて、個々の施策に反応する余地が生ずると考えられるからである。たとえば、フランス、スウェーデンの場合、長期に及ぶ家族政策の歴史の中で、人々は自らが受益者であることを経験を通して確信しているように見える。

ここで出生の質とは、たとえば全ての出生が親や社会から望まれたものであり、それゆえに出生後の心身の健全や良質な教育が保証された状態のことなどを指す。そのためには、妊娠・出産・子育てだけでなく、たとえば家庭や地域が安心、安全で親世代が将来に希望を持てるようなものでなくてはならないだろう。もちろん現在の「少子化対策」もこれらに貢献するものであるが、それらが誰のためになされているのかが不明瞭なときがあるように思う。たとえば、経済成長、社会保障制度維持や地域存続のための少子化対策（次世代の確保・健全育成）なのではなく、人々、とりわけ若年層や次世代の well - being 確保のための少子化対策であり、経済、社会保障、地域の維持発展であるということがもっと明瞭であってよい。

少子化対策を行う側は、用いる指標の特性や見方について正しく理解していることはもとより、出生を質として捉え、当事者達との継続的な対話などを通して、かれらのライフコースに寄り添って行く姿勢が求められるように思う。蛇足をいえば、フランスやスウェーデンの家族政策が実効性を有し、実質的出生率が人口置換水準付近で推移していることの背景には、両国の施策が第二次大戦直後以来のきわめて長期にわたる継続的で発展的な経過の中で、人々のライフコースに自然に寄り添うものになっているということがあるように思われる。

別出生の指標、働き方に関する指標などの分析は、こうした努力の例である。

マクロデータとミクロデータの統計的データ融合について

慶應義塾大学　経済学部・大学院経済学研究科
星野　崇宏
bayesian@jasmine.ocn.ne.jp

問題意識

政府統計やオープンデータの利活用の重要性

⇒特に基幹調査は報告義務を有することから代表性が担保されているものと考えることができる

＊調査の誤記入や学習効果などは別途考慮すべき

可能であれば個票データ（家計・個人・企業単位）を公開し利用

一方匿名性の担保は取得情報が豊富になるほど困難

⇒一般には集計情報として公開

民間としてはマーケティングや経営意思決定のために

代表性のある集計情報をいかに活用するかは重要な課題

⇒“ミクロデータとマクロ情報の統計的融合”へのニーズ

内容

- 統計的データ融合のレビュー
- 特にミクロデータとマクロデータの統計的データ融合
- 時系列データに対する統計的データ融合と選択バイアスの除去
- 準ベイズ推定によるミクロデータとマクロデータの融合解析
- 適用例

統計的データ融合とは？

複数の異なる情報源から得られたデータを統合し統計的推測を行う方法論

統計学／マーケティング分野では

Data fusion(Kamakura and Wagner,1997;Gillula and McCulloch,2013)

計量経済学では

Data Combination(Imbens and Lancaster,1994;Ridder and Moffitt,2008; Fan et.al,2014)

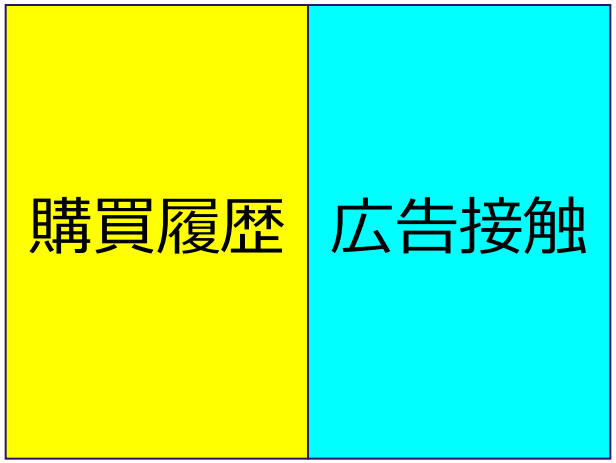
特に観測単位（以降ユニット：消費者／企業／地域など）が異なる複数のデータを活用して推論を行う方法

(c.f.)擬似パネル pseudo panel

シングルソースデータとマルチソースデータ

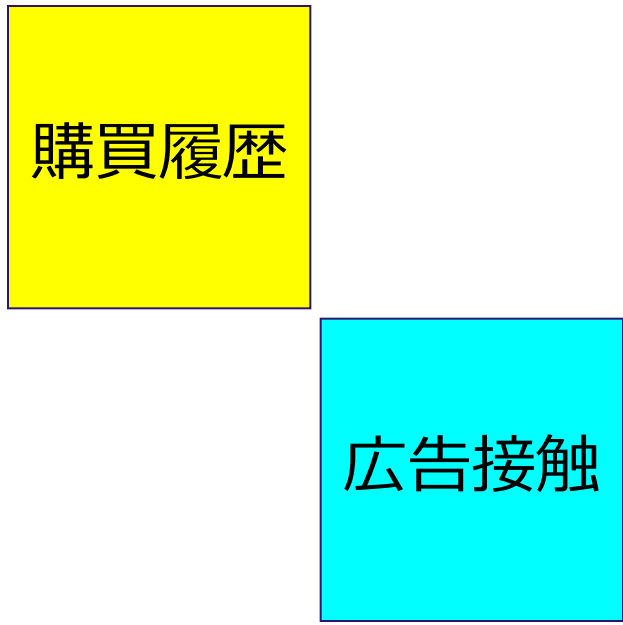
シングルソースデータ

自分の関心のある変数
すべてが、同じ対象者から
得られているデータ
⇒関連(例：広告効果)が分かる

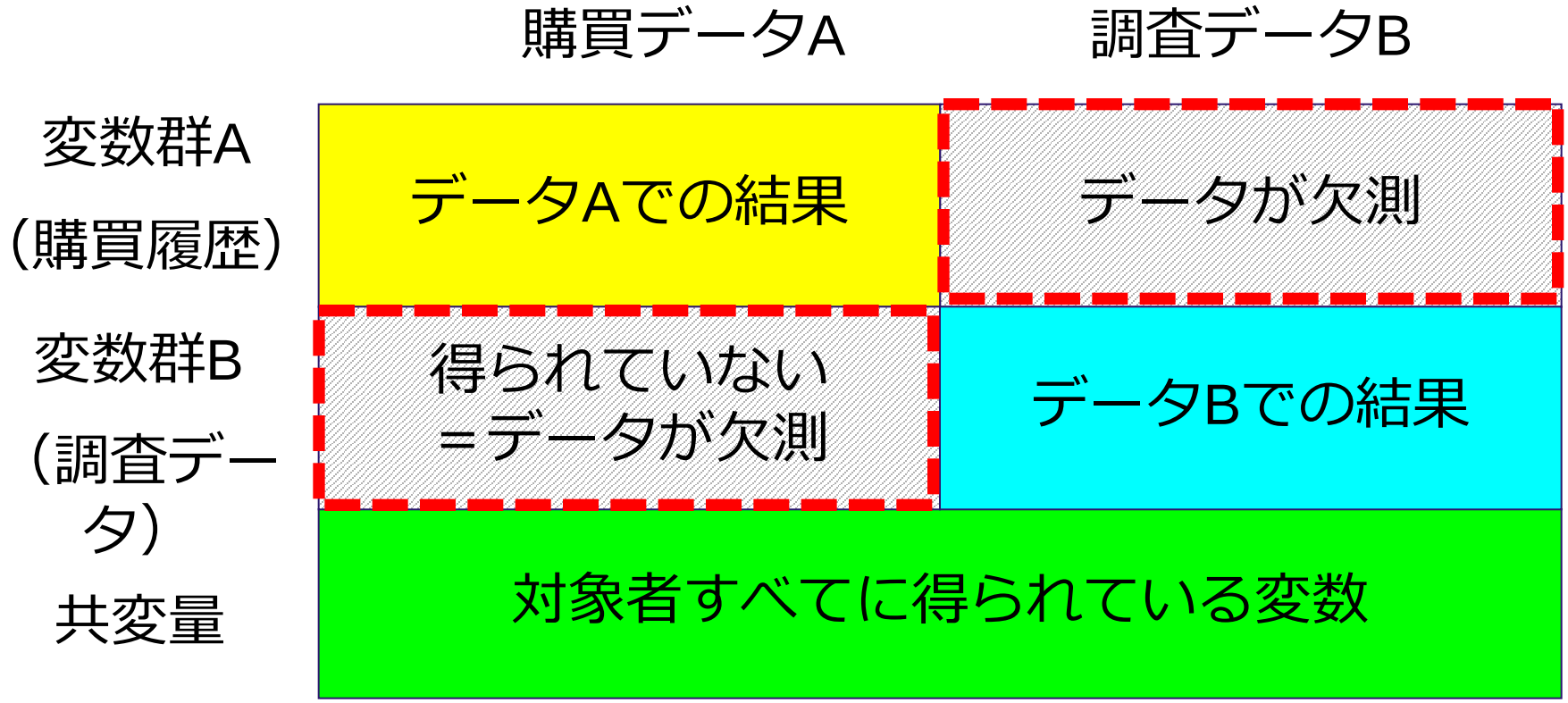


マルチソースデータ

自分の関心のある変数が
別々の対象者から分割して
得られているデータ
⇒普通はこれらからは関連は分からない



データ融合



ユニットが共通しない２つのデータから

・「変数Aと変数Bの関係（相関や回帰）」の推定

・欠測値の補完によるシングルソースデータ化

を行うことが目的とされる

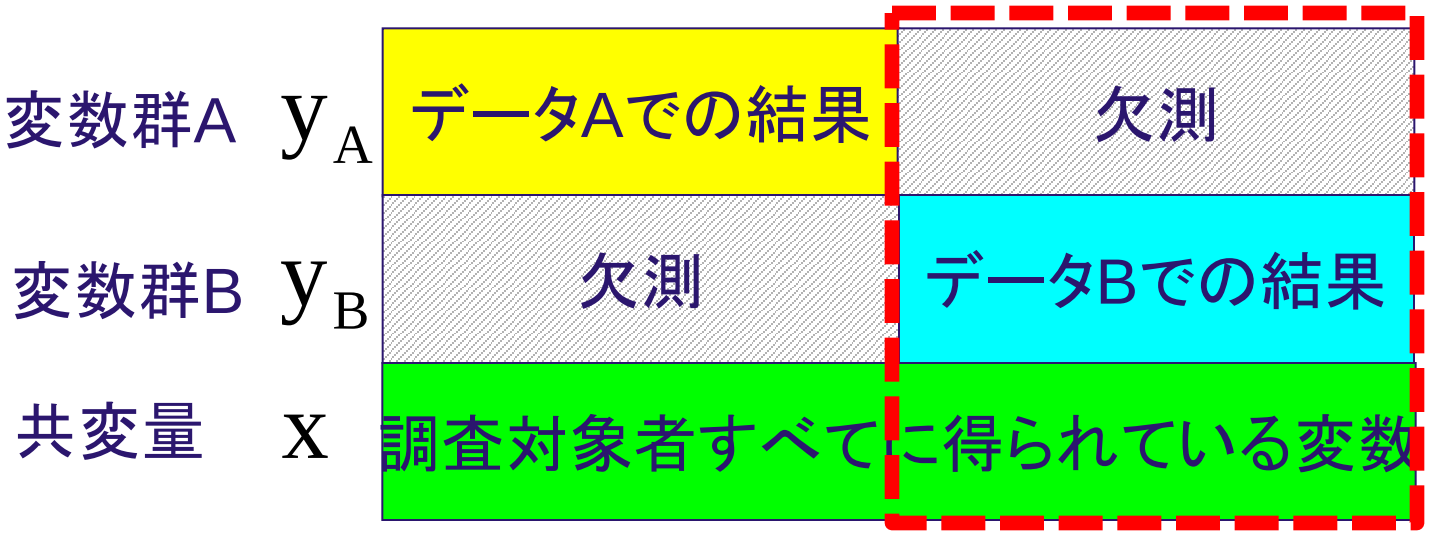
共変量情報を用いて データ融合が可能となる条件

y_Aの条件付き分布は？（データBでの）

$$p(y_A \mid y_B, z=0, x) = \frac{p(z=0 \mid y_A, y_B, x)\, p(y_A \mid y_B, x)}{p(z=0 \mid y_B, x)}$$

これは推定できない（ $z=0$ では y_A 欠測）

データA $z=1$　　データB $z=0$



共変量情報を用いて データ融合が可能となる条件

そこで　$p(z=0 \mid y_A, y_B, x) = p(z=0 \mid y_B, x)$

＝「ランダムな欠測」ならば

$$p(y_A \mid y_B, z=0, x) = p(y_A \mid y_B, x)$$

さらに「条件付き独立」

$$p(y_A, y_B \mid x) = p(y_A \mid x)\, p(y_B \mid x)$$

ならば

$$p(y_A \mid y_B, z=0, x) = p(y_A \mid x)$$

従って条件は

【1】「ランダムな欠測」である

【2】　y_A と y_B が条件付き独立である

2つの条件さえ成立すれば

欠測データがある場合の完全尤度

$$\prod_{i:z_i=1}^{N}\int p(y_{iA}, y_{iB} \mid x_i)p(z_i \mid y_{iA}, y_{iB}, x_i)dy_{iB}$$

$$\times\prod_{i:z_i=0}^{N}\int p(y_{iA}, y_{iB} \mid x_i)p(z_i \mid y_{iA}, y_{iB}, x_i)dy_{iA}$$

成立なら

$$=\prod_{i:z_i=1}^{N}\int p(y_{iA} \mid x_i)p(y_{iB} \mid x_i)p(z_i \mid x_i)dy_{iB}$$

$$\times\prod_{i:z_i=0}^{N}\int p(y_{iA} \mid x_i)p(y_{iB} \mid x_i)p(z_i \mid x_i)dy_{iA}$$

$$=\prod_{i:z_i=1}^{N}p(y_{iA} \mid x_i)p(z_i \mid x_i)\times\prod_{i:z_i=0}^{N}p(y_{iB} \mid x_i)p(z_i \mid x_i)$$

観測値だけから推定が可能

擬似パネル

経済学でよく利用される方法論(Hsiao,2003)

クロスセクションデータから「本来パネルデータでしかわからない結果を得る」

例）定年後の就業行動（岩本,2000）

定年後まで働きたいかどうか？には大きな個人差

説明変数：生年、学歴など

⇒説明変数以外の個人差は無視して解析

＊性年齢居住地域等で集計したデータを用いる場合もある

疑似パネル(Pseudo Panel)

	2000年の調査	2005年の調査
2000年での収入	2000年の調査結果	欠測
2005年での収入	欠測	2005年の調査結果
共通項目	調査対象者すべてに得られている変数	

2つの方法

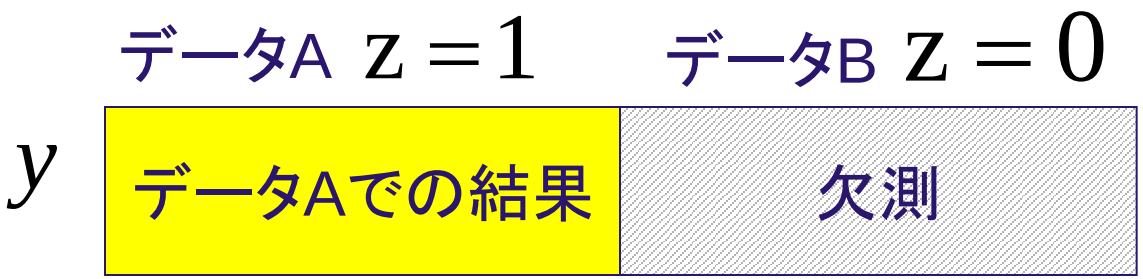
【統計的マッチング】他のデータ融合手法に同じ

【モデル仮定】例「性学歴生年コーホート内で個人差なし」として推定

Browning,Deaton and Irish,1985; Moffitt,1993；Hsiao,2003

マクロデータとの融合について

ミクロデータに加えてマクロデータがある場合は？



ケース1 ＋ データBでのy の平均等

または

ケース2 ＋ 全体でのyの平均等モーメント

＊上記標本全体（データAとB）で無作為抽出と仮定

ケース1）データAが無作為抽出である

ケース2）データAは無作為抽出ではない

＊通常はケース2（モーメントだけ代表性ある調査から）

Imbensらの一連の方法

【先ほどのケース1を仮定する場合】

（ミクロデータからのモーメント条件）

　通常の不偏推定方程式から誘導されるモーメント

（マクロデータからのモーメント条件）

　マクロデータにより得られる周辺分布のモーメント条件

例）プロビットモデル　サンプルサイズN

失業率を属性ベクトルxで説明　$P(y_i = 1 | x_i) = \Phi(x_i^t \beta)$

ミクロ：xの次元+1次元　$\sum_{i=1}^{N} \frac{y_i - \Phi(x_i^t \beta)}{\Phi(x_i^t \beta)(1 - \Phi(x_i^t \beta))} \phi(x_i^t \beta) x_i = 0$

マクロ：離散化した年代(j=1～J)ごとの失業率p_jの制約

$$\sum_{i:agec \cdot j=1} \Phi(x_i^t \beta) / \sum_{i=1}^{N} agec \cdot j - p_j = 0 \quad (j = 1 \cdots J)$$

＊マクロデータにおける周辺分布推定時の推定誤差も考慮可能

Imbensらの一連の方法

一般化モーメント法(GMM)による推定　具体的には

$$h(y_i, x_i \mid \theta) = \begin{pmatrix} h_1(y_i, x_i) \\ h_2(y_i, x_i) \end{pmatrix} = \begin{pmatrix} \dfrac{y_i - \Phi(x_i^t \beta)}{\Phi(x_i^t \beta)(1-\Phi(x_i^t \beta))} \phi(x_i^t \beta) x_i \\ z_{ij} \Phi(x_i^t \beta) / \sum_{i=1}^{N} z_{ij} - p_j \end{pmatrix} \quad g(y, x \mid \theta) = \frac{1}{N} \sum_{i=1}^{N} h(y_i, x_i \mid \theta)$$

として

$$\hat{\theta}_{GMM} = \arg\min_{\theta} g(y, x \mid \theta)^t W g(y, x \mid \theta)$$

・一般化線形モデルでの経験尤度法を用いた推定(Chaudhuriら 2008;JRSS,ser.B)

【先ほどのケース2を仮定する場合】

・同時分布のモーメント制約をもつ重み付き最小二乗法(Hellerstein and Imbens,1999)　・経験尤度法(Qin and Zhang,2007)

・Calibration estimator　・傾向スコア重み付け推定方程式　など

⇒方法論の仮定としてselection on observables(またはIgnorability)

ミクロデータに所属するかどうかが観測変数に依存

Calibration estimator

ケース2を仮定する方法のうち最もよく利用されている枠組み

事後層化(post-stratification)・Raking・(一般化)回帰推定を含む

1：すべての補助変数の重み付き集計値が（周辺）母集団集計値に一致するように「重み」を決める

2：それだけでは重み決定の自由度があるので、「抽出ウェイト」に”近く”なるように重みを決める

3：「近さ」を測る距離関数によって推定量が異なる

距離関数が線形関数(の2乗)の場合＝回帰推定量

ロジット関数の場合の特殊な例

＝傾向スコアによる重み付けと形式的に一致

＊但しその係数は補助変数の周辺情報によって決定される

問題点：ケース1の場合に利用できない／推定量の性質が不明確

周辺分布推定時の誤差を考慮できない

先行研究の問題点

通常マクロ情報とミクロデータを融合させるニーズはケース2

例）マクロ統計情報は全数調査ないしは偏りのないサンプリングによる調査（政府調査）

ミクロデータは偏りあるサンプリングでユニット抽出（企業データ）

利用するデータセット間でのユニットの抽出方法の差異（ケース2）を考慮した手法はいくつかの点で問題

具体的には

1）selection on unobservablesまたはmissing not at randomの状況では推定にバイアス

2）回帰モデルで説明変数にのみ抽出が依存するという仮定(MAR)ならばそもそもweightによる調整は不要

抽出の偏りについて

無作為抽出標本からのミクロデータの偏りを考えるために以下の欠測データの枠組みを考える

Selection on observable /Ignorability/MARの仮定は

y

データA $z=1$　　データB $z=0$

ミクロデータ	欠測

$$p(z=1|y,x)=p(z=1|x)$$

ここで回帰モデル $E(y|x,\beta)$ の母数 β の推定に際して不偏推定方程式 $\sum_{i=1}^{N} m(y_i|x_i,\beta)=0 \quad s.t. \quad E_{y|x}(m(y|x,\beta))=0$ はデータAだけでも不偏

$$E(z\times m(y|x,\beta))=E_{z|x}(z)E_{y|x}(m(y|x,\beta))=0$$

⇒抽出がアウトカムには依存しない場合にはマクロ情報を利用した重み付けはあまり意味がない

古典的な方法としてのHeckman選択モデル

アウトカムyが観測されるメカニズムを考慮

Outcome equation $y_i = x_i^t \beta + u_i$ $y_{is} > 0$ なら y_i が観測

Selection equation $y_{is} = x_i^t \beta_s + u_{is}$

同時分布を仮定

$$\begin{pmatrix} y \\ y_s \end{pmatrix} \sim N\left(\begin{pmatrix} x_i^t \beta \\ x_i^t \beta_s \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

このときyが観測される確率は

$$p(y_s \le 0 \mid y, x) = 1 - \Phi\left(\frac{1}{\sqrt{1-\rho^2}} \left\{ x_i^t \beta_s + \frac{\rho}{\sigma}\left(y - x_i^t \beta\right) \right\} \right)$$

⇒MARではない(y_s<0でunobservableであるyに依存)

推定の識別性問題(特にρ)／分布仮定からの逸脱に敏感

Little(1983), Copas and Li(1997)

⇒実際の解析ではあまり利用されない

新しい解析手法
GMMの準ベイズ推定

新手法の目的

「ミクロデータ」と「マクロ情報」の融合において
先行研究で考慮されなかった以下の点を解決
・ミクロデータがMNARであることを考慮することができる
・マクロ情報があることで識別性が担保される
・マクロ情報を推定する際の推定誤差も考慮する

具体的な適用の想定例としては
（ミクロデータ）企業の保有する購買履歴データ
（マクロデータ）家計調査の集計データ
⇒ミクロデータは特定の小売りチェーンで購入しているので“偏りを有する”データ
例えば
自社以外での購買と自社で得られる情報の関係の理解

MNARのモデルについて

モデルの尤度は以下のように表現される。

$$\prod_{i=1} p(y_i \mid x_i, \theta)^{z_i} \left[\int p(y_i \mid x_i, \theta)\, p(z_i \mid y_i, x_i, \phi) dy_i \right]^{1-z_i}$$

上記尤度から構成される不偏推定方程式は下記

$$g_1(y, x \mid \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{N} z_i \log p(y_i \mid x_i, \theta) + (1 - z_i) \log \left[\int p(y_i \mid x_i, \theta)\, p(z_i \mid y_i, x_i, \theta) dy_i \right] = 0$$

または

$$g_1(y, x \mid \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{N} \frac{z_i}{p(z_i \mid y_i, x_i)} \log p(y_i \mid x_i, \theta) = 0$$

但し一般的には上記のモデルの識別性のために

補助情報としてマクロ情報が与えられるとする。

$$g_2(y, x \mid \theta) = m(y, x \mid \theta) = 0$$

＊具体的形式はモデルと情報に依存：yの平均や分散など

合わせるとモーメント制約は

$$g(y, x \mid \theta) = \begin{pmatrix} g_1(y, x \mid \theta) \\ g_2(y, x \mid \theta) \end{pmatrix} = 0$$

準ベイズ推定

スコア関数の部分で欠測値についての積分が必要

本研究では擬ベイズ推定法に基づくマルコフ連鎖モンテカルロ法を用いた推定を利用する。

【準ベイズ推定】quasi-Bayesian estimation using MCMC

Chernozhukov and Hong (2003)

通常の事後分布 $$p(\theta \mid y) = \frac{p(\theta)\exp l(\theta \mid y)}{\int p(\theta)\exp l(\theta \mid y)d\theta}$$

但し $l(\theta \mid y)$ は対数尤度

ではなく擬似事後分布

$$q(\theta \mid y) = \frac{p(\theta)\exp l^q(\theta \mid y)}{\int p(\theta)\exp l^q(\theta \mid y)d\theta}$$

但し $-2l^q(\theta \mid y)$ はGMMの基準関数

とすると擬似事後平均は一致性があり漸近正規性を有する

準ベイズ推定によるミクロマクロ融合

【準ベイズ推定の利点】

擬似事後分布 $q(\theta \mid y)$ からの乱数列を利用して

マルコフ連鎖モンテカルロ法(MCMC)により統計的な推測が可能

⇒欠測値もMCMCの途中で擬似予測事後分布から発生すればよい

【基準関数】(Igari and Hoshino,2015;Hoshino,2016)

今回の状況ではミクロとマクロで独立な情報のため

$$l^q(\theta \mid y) = \left(g_1^t(y,x \mid \theta)\ g_2^t(y,x \mid \theta)\right)\begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}\begin{pmatrix} g_1(y,x \mid \theta) \\ g_2(y,x \mid \theta) \end{pmatrix}$$

$$= g_1^t(y,x \mid \theta)W_1 g_1(y,x \mid \theta) + g_2^t(y,x \mid \theta)W_2 g_2(y,x \mid \theta)$$

＊注意：ベイズでは事前知識を事前分布で表現できるとされるが、
一般に母数の関数についての情報は明示的に表現できない

具体的なモデルの例

（1）標本調査における未回収バイアスの補正

回収有無がモデリングの対象となる変数yやxに依存

【マクロ情報】YやXの周辺分布、一部変数での同時分布

（2）競合他社での購買金額の推定

$y = y_I + y_C$: y_Iは自社での購入額 y_Cは他社での購入額

自社が持っている情報はYIと個人属性・来店履歴等説明変数X

$p(y \mid x, \theta)$ または $p(y_C \mid x, \theta)$

の母数推定／予測分布の推定に関心がある

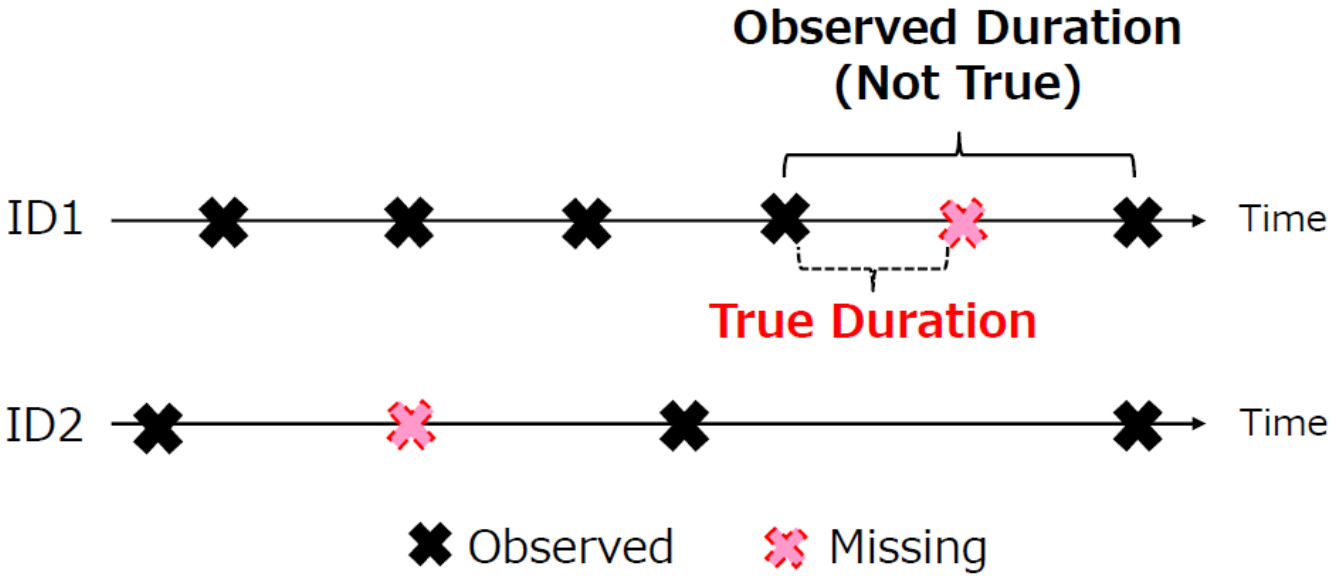
【マクロ情報】Yの周辺分布、あるいはXの一部とYの集計表

⇒観測されない潜在変数 $y_C = y - y_I$ についてyの周辺分布のモーメント情報とy_I, Xのモーメント情報から推論可能

具体的なモデルの例

（3）真の購入頻度の推定(Igari and Hoshino,2015)

他社でも当該商品
購入
⇒自社の情報だけ
では真の間隔は
不明
繰り返しのある
Duration analysis



属性や価格、企業のプロモーション変数で説明したい

【マクロ情報】

家計調査での当該商品の属性別年間購入数量のクロス集計表

（4） 集計レベルの時系列データと個人・家計レベルの個票データの融合：状態空間モデルでの時変係数を仮定を置かず推定

まとめと議論

- シミュレーションの結果と具体的な解析例については当日報告
- 今回は公的統計を民間がいかに利活用するかという観点から

“ミクロデータとマクロ情報の統計的融合”の方法論の紹介と新しい解析手法の枠組みについて紹介

- 但し誤記入や学習効果の可能性がある公的調査の“バイアス補正”にも活用可能と考えられる

例）家計レベルのミクロデータ「家計調査」と売上ベース・小売業による集計された地域別品目別の「商業動態統計調査」の融合

⇒周辺分布は「商業動態統計調査」に従う「家計調査」個票データ

- 公的調査の匿名データが利用可能になった場合には

“マルチソースのミクロデータの統計的データ融合”の方法論のさらなる発展と利用が望まれる

Illustrating Evidence Based Anonymization

Nobuaki HOSHINO*

January 2016

Abstract

To publish a data set we must ensure that included individuals are unidentifiable. Nevertheless this unidentifiability is often judged rather subjectively. The author has proposed objectifying this practical judge by the statistical estimation of a model, where an unknown parameter describes uncertainty on identification. This idea exploits observation as a statistical evidence. To promote this evidence based method, the present article demonstrates its applicability to the decision of the acceptable range of disclosure risk on Japanese Anonymized Data.

Key words: Population unique, Privacy, Statistical Disclosure Control.

1 Introduction

Confidentiality or privacy reasons let laws allow no individual to be identified when data are published. Many techniques to prevent such identification are known as anonymization; see e.g. Willenborg and de Waal (1996, 2000). After anonymizing a data set, data publishers need to decide whether the set is identifiable or not. This decision, however, has been made in a rather subjective manner.

Technical efforts to formalize this decision result in many measures of (re-)identification risk, but no objective method to decide a threshold of those measures seems known. This is so because those researches authorize this decision to depend on a personal preference under a tradeoff between re-identification risk and utility of data; see e.g. Duncan et al. (2001).

However, identifiability does not depend on a personal preference: When data are identifiable they are so regardless of the preference of their publisher. Also laws say nothing about the utility of data. In other words, no identifiable data set is publishable even if they are highly useful. Therefore in practice the true goal of anonymization is not to take the balance of risk and utility but to maximize utility within an acceptable range of risk. This goal has been claimed by Dalenius (1977) and others, and for a practitioner the decision of the threshold of identification risk is the primary issue.

This important decision should be objective and explicable. To achieve this clarity the present article employs Hoshino's (2013) statistical model of identification with an unknown threshold of risk, and estimates it by observing whether published data have been identified

*School of Economics, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

or not. These observations are statistical evidence that carries information on the threshold of identification risk. Thus we call our approach "Evidence Based Anonymization (EBA)".

In this way an estimated threshold depends on the measured elements of risk. Our risk measure only accounts for the difference of data and neglect institutional effects as typical risk measures do. Institutional effects, however, seem to exist since the degree of anonymization empirically depends on the qualifications of users of data: A public use file tends to be more anonymized than a scientific use file. Therefore in order to control institutional effects that may exist, we restrict our sample space to the cases of each institution. In other words, the threshold of our risk measure can differ among institutions. Such a difference arises from the difference of pooled institutional effects.

Restricting our sample space is advantageous in that we do not have to decompose institutional effects. This decomposition is virtually impossible due to very limited information on the latent ability of identification that a society possesses.

To exemplify EBA, the present article estimates the risk threshold of Japanese Anonymized Data of surveys conducted by Ministry of Internal Affairs and Communications (MIAC). Anonymized Data are defined by Statistics Act so that no individual shall be identified. Global recoding and record suppression are mainly used to satisfy this definition.

MIAC's Anonymized Data have been available on 4 surveys (Employment Status Survey, Housing and Land Survey, National Survey of Family Income and Expenditure, and Survey on Time Use and Leisure Activities) from 2009, on Labour Force Survey from 2011, and on Population Census from 2013. The Japanese society has not recognized any identification of an individual contained in these data as of 2015.

Anonymized Data are provided for academic research and advanced education under a license. The number of past users of these data varies among surveys, but in total about 30 applicants have passed the same review procedure for use in each year as of 2013. We consider that these cases of using Anonymized Data share the same threshold of identification risk, since they are under the same institutional measures against identification.

The present article is organized as follows. Section 2 explains our statistical model that links our risk measure to the observation of identification. Section 3 explains the detail of evaluating our risk measure. Section 4 estimates the threshold of identification risk of Anonymized Data. Section 5 concludes.

2 Statistical model of identification

This section explains Hoshino's (2013) method to decide whether a measured identification risk is acceptable or not. The first subsection technically describes unidentifiability. The second subsection presents a model whose parameter expresses uncertainty on identification. The third subsection statistically estimates this paramter, by which the acceptable range of our risk measure is determined.

2.1 Definition of identifiability

An effort on modeling identification can be seen in Marsh et al. (1991). They argue that the probability of identification is the product of the following probabilities:

$$\Pr(\text{actual identification}) = \Pr(\text{success of identification}|\text{trial of identification}) \Pr(\text{trial of identification}). \quad (1)$$

In eq. (1), the event of "actual identification" is regarded as the joint event of "success of identification" and "trial of identification". This discrimination between "actual identification" and "success of identification" corresponds to different legal concepts of anonymity or unidentifiability.

Absolute anonymity, using a German legal term, is a state where the possibility of identification is eliminated with no doubt. We regard this state as equivalent to a state that

$$\Pr(\text{success of identification}|\text{trial of identification}) = 0. \quad (2)$$

De facto anonymity, which is also a German legal term, is a state where the cost of identification dominates the benefit of identification. In this case the probability of the trial of identification should be low, and thus we regard this state as equivalent to a state that Pr(actual identification) is low.

The present article focuses upon the assessment of the absolute anonymity, because Japan Law seems to define confidentiality as such. Consequently we evaluate whether the conditional probability of "success of identification" given "trial of identification" is zero or not.

Marsh et al. (1991) regard the success of identification as the result of matching betweeen a published file and an outer file owned by an attacker (who tries to identify an individual). They accordingly propose the following factorization:

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c), \quad (3)$$

where the events from a to d are

(a) On the attribute of the same individual, both a published file and an outer file record the same value (i.e. no misclassification etc.).

(b) A published file contains an individual.

(c) An individual is a population unique.

(d) A population unique is verified to be so.

If we can evaluate the probabilities of the right hand side of eq. (3), we can obtain the conditional probability of "success of identification" given "trial of identification". However, the evaluation of these probabilities by Marsh et al. (1991) is not convincing from a modern point of view.

As discussed in Section 3, $\Pr(a,b,c)$ depends on information that an attacker currently knows, yet $\Pr(d|a,b,c)$ should depend on information that an attacker does not currently know. An attacker may be able to collect additional information to verify a population unique. Such currently nonexistent information is unobservable and hard to estimate. Hence the author considers that no one can plausibly evaluate $\Pr(d|a,b,c)$.

Now let us be reminded that we just would like to know whether eq. (2) holds or not. This evaluation is far easier than to evaluate the conditional probability of "success of identification" given "trial of identification".

Therefore we rewrite eq. (3) as

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a,b,c)\Pr(d|a,b,c). \tag{4}$$

Then we can see that eq. (2) holds if and only if at least one of $\Pr(a,b,c)$ and $\Pr(d|a,b,c)$ is zero. On data for scientific purposes, $\Pr(a,b,c)$ is usually positive. Consequently our usual assessment on unidentifiability reduces to a decision whether $\Pr(d|a,b,c)$ equals zero or not. Since the direct evaluation of $\Pr(d|a,b,c)$ is hopeless, we will estimate whether $\Pr(d|a,b,c)$ equals zero or not.

2.2 Model for discerning identifiability

From our argument so far, we would like to discern whether

$$\Pr(d|a,b,c) = 0 \tag{5}$$

or not, since eq. (5) is sufficient for an unidentifiable state.

To this goal we note that $\Pr(d|a,b,c)$ is subject to the event of (a,b,c), and $\Pr(a,b,c)$ can be evaluated since it only depends on existent information. The increment of $\Pr(a,b,c)$ implies that more information about population uniques is published. The more information exists, more easier the verification of a population unique should become. Hence the conditional probability of d given (a,b,c) should be monotonically increasing as $\Pr(a,b,c)$ increases. If so, there exists nonnegative β such that

$$\Pr(a,b,c) \leq \beta \Leftrightarrow \Pr(d|a,b,c) = 0. \tag{6}$$

Then we conclude that the assessment of identifiability reduces to the evaluation of $\Pr(a,b,c)$, since eq. (2) is tantamount to $\Pr(a,b,c) = 0$ or $\Pr(d|a,b,c) = 0$.

In the model (6), $\Pr(a,b,c)$ can be interpreted as the easiness of identification. This is a type of re-identification risk measure, and its threhold β is unknown. We decide it by statistical estimation in the following.

2.3 Observational model of identification

For the statistical estimation of β in eq. (6), we need an observation that carries information on β. Hence we would like to observe the event of d or the success of identification. However, a society may not always recognize such an event; a successful attacker may hide. Therefore we discriminate "actual identification" from its social recognition.

Let a random variable X be 1 when "actual identification" is socially recognized, and 0 otherwise. That is,

$$\Pr(X=1) = \Pr(\text{recognized}|\text{actual identification})\Pr(\text{actual identification}).$$

Then from eq. (1) and eq. (3),

$$\begin{aligned}\Pr(X=1) &= \Pr(\text{recognized}|\text{actual identification}) \\ &\quad \times \Pr(a,b,c,d)\Pr(\text{trial of identification}).\end{aligned} \tag{7}$$

Further, let us write the evaluated value of $\Pr(a,b,c)$ as γ, and write

$$p(\gamma) = \gamma\Pr(d|a,b,c)\Pr(\text{recognized}|\text{actual identification})\Pr(\text{trial of identification}). \tag{8}$$

Then

$$\Pr(X=1) = \begin{cases} p(\gamma) & \text{if } \gamma > \beta \\ 0 & \text{if } \gamma \leq \beta. \end{cases} \tag{9}$$

If $p(\gamma)$ is positive, the observed value of X carries information on β, and we can estimate β from X's. Actually $p(\gamma)$ is positive when both

$$\Pr(\text{recognized}|\text{actual identification})) > 0 \tag{10}$$

and

$$\Pr(\text{trial of identification}) > 0 \tag{11}$$

hold. The first condition (10) should be satisfied because an attacker has an incentive to show off their success of identification. Also hiding through is not always possible. The second condition (11) should also be satisfied because of a potential incentive to do so. Hence we regard that $p(\gamma)$ is positive. It is worth mentioning that we assume no specific form of $p(\cdot)$.

Suppose that there are n past experiences of publishing anonymized data. We regard these as independent samples from the model (9). For the i-th, $i = 1, 2, \ldots, n$, sample we measure $\Pr(a, b, c) = \gamma_i$ and observe the social recognition of actual identification $X_i = x_i$. Write the likelihood of the observations as $\ell(\beta)$. To simplify our argument we assume that $\gamma_1 > \gamma_2 > \cdots > \gamma_n$.

Now we consider the maximum likelihood estimator $\hat{\beta}$ of the threshold. If there exists an integer m such that $x_{m-1} = 1, x_m = x_{m+1} = \cdots = x_n = 0$, then $\ell(\beta) = 0$ for $\beta \geq \gamma_{m-1}$, $\ell(\beta) \propto p(\gamma_{m-1})$ for $\gamma_{m-1} > \beta \geq \gamma_m$, and $\ell(\beta) \propto p(\gamma_{m-1}) \prod_{j=m}^{i} (1 - p(\gamma_j))$ for $\gamma_i > \beta \geq \gamma_{i+1}, i \geq m$. Hence $\gamma_{m-1} > \hat{\beta} \geq \gamma_m$ because $p(\gamma)$ is positive. If there exists no social recognition of actual identification, then $\hat{\beta} \geq \gamma_1$.

In general we denote the lowest easiness of identification among samples with social recognition of actual identification by γ^-. If there has been no such recognition, let γ^- be 1. Also among samples with the easiness that is lower than γ^- we denote the highest easiness of identification by γ^+. Then $\gamma^+ \leq \hat{\beta} < \gamma^-$.

3 Measuring disclosure risk

To substantiate our theoretical model of the previous section, we have to establish the method of evaluating $\Pr(a, b, c)$, which is explained in this section following Hoshino (2013). The first subsection clarifies the policy of selecting key variables or quasi-identifiers. Under this policy the second subsection decomposes $\Pr(a, b, c)$ to the product of $\Pr(a), \Pr(b|a)$ and $\Pr(c|a, b)$, and explains the method of evaluating each probability.

3.1 How to select key variables

First we discuss the policy of selecting key variables for matching, on which the number of population uniques heavily depends. Because EBA compares $\Pr(a, b, c)$ among cases, the policy must be fixed.

Few arguments exist on the formal selection of key variables. For example, Elliot et al. (2010, 2011) claim a comprehensive survey of existent information about individuals in a society for this purppose. The author never denies the importancd of such information, but their argument

does not directly result in the best selection of key variables. Fung et al. (2010) describe the selection of key variables as "an open problem".

The present article selects key variables to best estimate β. Existing researches can not optimize the selection because they do not consider the aftermath of evaluating population uniques.

Suppose that there are k variables in a published file. Then there are 2^k ways to select key variables in theory. The number of population uniques can be evaluated in each way, and we write the order statistics of these numbers as $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)}$. Then the issue of the selection of key variables is nothing but the selection of a rank among $(1, 2, \ldots, 2^k)$, over which attackers are distributed subject to their knowledge about individuals.

For given data we evaluate $\Pr(a, b, c)$ at a selected rank r, and compare it with γ^+, which is also evaluated under the same policy of selecting key variables. Suppose that $\Pr(a, b, c)$ at the r-th rank is smaller than γ^+. Then, for fixed $\Pr(a, b)$, the given data should be safe against attackers who lie on ranks smaller than r, since $u_{(i)} \leq u_{(r)}$ for $i \leq r$. The given data, nevertheless, have no evidence of safety against attackers who lie on ranks larger than r.

Thus one might think that we should select the largest rank: 2^k. However, an attacker may not exist on the 2^k-th rank. If so, an observed X of eq. (9) carries no information on the safety of $\Pr(a, b, c)$ at the 2^k-th. Hence, considering the distribution of attackers over the ranks, we should select the largest rank on which an attacker exists.

The best way to select key variables has been described theoretically, but in practice, the distribution of attackers is unknown. Therefore we have to estimate the maximum of the distribution of attackers over ranks. The precise estimation of a maximum is, however, known to be difficullt, the theory of extreme values might be usable though. Also an errorneous estimate of the maximum rank leads to unstable $\hat{\beta}$. Hence, as a second best way, we should estimate a percentile, which is less difficult. For example it should be more practical to estimate the 99th percentile of the distribution of attackers, as in the case of financial risk called Value at Risk (VaR).

Unfortunately the quantitative evaluation of such a percentile is virtually impossible since we can not observe the distribution of attackers. Hence we select key variables whose information is publicly known. This policy is common in practice, which actually implies that some large percentile is estimated.

Our policy sacrifices anonymization's grip on the strongest attackers, but other institutional protection may suffice. For example, since the strongest attackers should be conspicuous, a data publisher may be able to reject their request for a scientific use file. It censors the right tail of the distribution of attackers. A penal code should be effective even in the case of a public use file.

3.2 Risk measured to control most

This section describes the method of measuring $\Pr(a), \Pr(b|a)$ and $\Pr(c|a, b)$ under our policy of measuring risk at a large percentile.

3.2.1 Measuring $\Pr(a)$

The attribute of an individual may be differently recorded between a published file and an outer file. Marsh et al. (1991) ascribe this difference to an error in recording or a change of an

attribute with the passage of time. A perturbation technique such as swapping can also cause this difference.

If at least one key variable of an individual is affected by these causes, then the event of a does not occur. Therefore Marsh and others claim that the increment of key variables tends to decrease $\Pr(a)$; Shlomo and Skinner (2010) give a numerical example of this kind.

Nevertheless they neglect the possibility of the correction of such differences. A record-linkage-like technique can correct them especially when a unique individual lies in a sparse space. Because this sparsity emerges when key variables increase, the increment of key variables does not necessarily decrease $\Pr(a)$. Hence we do not relate $\Pr(a)$ to the number of key variables.

Consequently we evaluate $\Pr(a) = 1$ for an unperturbed file. This evaluation does not imply no error in recording. The rate of errors, which is uncontrolable by a data pulisher, is a part of uncertainty on identification. Hence we consider that the rate of errors too is described by β.

The effect of perturbation should be evaluated casewise. The present article does not deal with a perturbed file, and thus we do not argue further.

3.2.2 Measuring $\Pr(b|a)$

Following Marsh et al. (1991) we evaluate $\Pr(b|a)$ as the ratio of the size of a published file to the correspoinding population size.

3.2.3 Measuring $\Pr(c|a, b)$

Marsh et al. (1991) defines $\Pr(c|a, b)$ by a ratio of the number of population uniques to its population size. The evaluation of this ratio usually involves estimating the number of population uniques, which is not straightforward.

On this estimation we employ Hoshino's (2001) method that exploits Pitman's (1995) sampling formula. Our method is advantageous in that it does not require tailored modeling for each data set. It is thus suitable for comparing many data sets with the same standard.

Regression is a common way to estimate the number of population uniques, but as Skinner and Shlomo (2008) address, such an inference may be sensitive to the specification of a model. Therefore regression is unsuitable for our comparison of estimates.

Another advantage of our method is its applicability to sparse data. Many key variables are likely to be selected under our policy. Then individuals are distributed over the high dimensional space of these key variables, and a frequency on a location in this space tends to be zero, which implies sparse data. If most of these frequencies are zero, such a location indexed by the values of key variables has little information on the number of uniques. That is, key variables are useless to estimate the number of uniques: Regression fails. Our method still works.

4 Risk of Anonymized Data

This section demonstrates the estimation of β for the cases of Anonymized Data. Results are to be presented only orally.

5 Concluding remarks

The implication of our argument in Section 2 is clear: A given data set is publishable if its $\Pr(a, b, c)$ does not exceed γ^+, since it has a statistical evidence of unidentifiability. This γ^+ is estimated in Section 3 for Anonymized Data provided by MIAC.

By these arguments the present article illustrates one method to objectively decide whether given data are identifiable or not. Subjective decision may employ past experiences implicitly, yet our method explicitly employs past experiences as a statistical evidence.

What if there has been no past example that can be an evidence? Then begin with publishing apparently safe data; a clinical trial decides the threshold of some dose by gradually increasing risk. The idea of evidence based decision originates in medicine. It can also be applied to anonymization.

Acknowledgements

This research has been supported by Kakenhi grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Anonymized Data used in the present article have been provided by the National Statistics Center of Japan, subject to Statistics Act.

References

[1] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, **15**, 428–444.

[2] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[3] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[4] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[5] Fung, B.C.M., Wang, K., Fu, A.W.C and Yu, P.S. (2010) *Introduction to Privacy-Preserving Data Publishing*, CRC Press, New York.

[6] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment, *Journal of Official Statistics*, **17**, 499–520.

[7] Hoshino, N. (2013) *Evidence Based Anonymization*. Discussion Paper No.21, Faculty of Economics and Management, Kanazawa University. (In Japanese.)

[8] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.

[9] Pitman, J. (1995) Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[10] Shlomo, N. and Skinner, C. (2010) Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, **4**, 1291–1310.

[11] Skinner, C. and Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Log-Linear Models. *Journal of the American Statistical Association*, **103**, 989–1001.

[12] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice* , Lecture Notes in Statistics 111, Springer, New York.

[13] Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics 155, Springer, New York.

季節調整プログラム X-13ARIMA-SEATS について

高岡　慎

琉球大学法文学部

2016 年 1 月 29 日

科研費プロジェクト『経済統計・政府統計の理論と応用からの提言』カンファレンス

はじめに

■ 本報告の内容

1. X-11 から X-13ARIMA-SEATS へ
2. RegARIMA モデルの概要
3. TRAMO-SEATS による季節調整
4. X-13ARIMA-SEATS
5. 調整結果の比較
6. まとめ

1. X-11 から X-13ARIMA-SEATS へ

■ X-11

- Shiskin, Young, and Musgrave(1967)
- 移動平均フィルタの連続的な適用による時系列の分解
- 単純な対称移動平均フィルタ
- ヘンダーソン移動平均フィルタ
- 端点付近では非対称フィルタ（マスグレーブ法）

1. X-11 から X-13ARIMA-SEATS へ

■ X-11-ARIMA

- カナダセンサス局（Dagum(1988)）
- RegARIMA モデルの導入
  ⇒RegARIMA モデルによる予測値で時系列を延長
- 対称移動平均フィルタの適用

1. X-11 から X-13ARIMA-SEATS へ

■ X-12-ARIMA

- Findley, Monsell, Bell, Otto, Chen(1998)
- RegARIMA モデルの使用
  ⇒ 異常値・レベルシフトなどの変動を回帰変数として処理
- モデルによる系列の延長
- X-11 フィルタによる処理
- 事後診断機能など、安定した季節調整のための様々な改良

1. X-11 から X-13ARIMA-SEATS へ

■ TRAMO-SEATS

- スペイン銀行（Maravall(1995) その他）
- RegARIMA モデルの利用
  ⇒ 異常値・レベルシフトの処理と時系列構造の特定
- TRAMO パートでのモデル選択
  ⇒ 単位根検定と情報量規準による選択
- SEATS パートでの信号抽出に基づく季節調整
  ⇒WK(Wiener-Kolmogorov) フィルタによる時系列の分解

1. X-11 から X-13ARIMA-SEATS へ

■ X-13ARIMA-SEATS

- X-12-ARIMA と TRAMO-SEATS の統合
- TRAMO パートを RegARIMA モデルの自動モデル選択コマンドとして内蔵
- SEATS の機能を全て実装
- 季節調整処理では X-11 法と SEATS を選択可能

2. RegARIMA モデルの概要

■ RegARIMA モデル

原系列 y_t が

$$y_t = \sum_i \beta_i x_{it} + z_t$$

のような線形回帰モデルの形式で表され、z_t が季節 ARIMA モデルに従うとき、y_t のモデルを RegARIMA モデルとよぶ。一般的な表記は

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\left(y_t - \sum_i \beta_i x_{it}\right) = \theta(B)\Theta(B^s)a_t$$

となる。B はバックシフトオペレータ、$\phi(B), \Phi(B^s), \theta(B), \Theta(B^s)$ は B の多項式、a_t はホワイトノイズ。

2. RegARIMA モデルの概要

■ モデルの選択

- **pickmdl** コマンドによりモデルの次数を自動選択することが可能。
- **pickmdl** コマンドは X-12-ARIMA では **automdl** という名称だったが、X-12-ARIMA の最終バージョンでは TRAMO のモデル選択法 (後述) が導入され、従来の選択アルゴリズムは **pickmdl** に名称が変更された。
- 候補となるモデルのインサンプルでの backcast error と forecast error が計算され、一定の規準を満たしているモデルが選択される。
- 階差次数も同時に選択するために、予測誤差に基づく経験的な統計量が採用されている。
- 候補が全てリジェクトされる場合もあり、使いにくい点があった。

2. RegARIMA モデルの概要

■ モデルの推定

繰り返し一般化最小二乗法 (IGLS) による推定

(1) 原系列と説明変数系列の両方に必要な階差を適用

(2) 与えられた AR と MA 母数に対して回帰係数を一般化最小二乗法 (GLS) で推定

(3) 回帰モデルの母数 β_i の値を所与として最尤法により ARMA モデルの係数を推定

(4) (2) と (3) のプロセスを収束するまで反復

3. TRAMO-SEATS による季節調整

■ TRAMO-SEATS の概要 1

- スペイン銀行により開発されたモデルベースの季節調整法 (Gomez and Maravall(1996))
- TRAMO(Time series Regression with ARIMA noise, Missing Observation and Outliers) パートと SEATS(Signal Extraction in ARIMA Time Series) パートに分かれており、前者では原系列に適用すべき RegARIMA モデルの選択・推定、後者では信号抽出による季節調整に関する処理が行われる。

3. TRAMO-SEATS による季節調整

■ TRAMO-SEATS の概要 2

- RegARIMA モデルは

TRAMO パートの事前調整
- RegARIMA モデルの適用
- 外れ値の処理
- 欠損値の処理
- 前方予測と後方予測の追加

などの処理に利用される。

- 事前調整が行われた後で、WK フィルタによる時系列の分解が行われる。
- 全体の処理の流れは X-12-ARIMA のプロセス類似しているが、利用するフィルタが異なる。

3. TRAMO-SEATS による季節調整

■ 信号抽出による時系列の分解 1

系列 $\{X_t\}$ がシグナル (S_t) とノイズ (N_t) の和

$$X_t = S_t + N_t$$

であるとする。シグナルとノイズはいずれも定常で、かつ互いに独立。
$\{X_t\}$ のみが観察可能であるとき、X_t を利用して S_t を推定する。
線型な推定量

$$\hat{S}_t = \sum_{j=-\infty}^{\infty} \varphi_j X_{t-j} = \varphi(B) X_t$$

を考え、最適なウェイトを考える。

3. TRAMO-SEATS による季節調整

■ 信号抽出による時系列の分解 2

平均 2 乗誤差

$$E\left[(S_t - \hat{S}_t)^2\right]$$

を最小にするものを最適なウェイトと考えると、

$$\hat{S}_t \quad = \quad \varphi(B) X_t = \frac{\sigma_s^2}{\sigma^2} \frac{\psi_s(B)\psi_s(B^{-1})}{\psi(B)\psi(B^{-1})} X_t$$

が得られる。
⇒WK(Wiener-Kolmogorov) フィルタ

3. TRAMO-SEATS による季節調整

■ 信号抽出による時系列の分解 3

ただし

$$\begin{aligned} X_t &= \psi(B)\epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2) \\ S_t &= \psi_s(B)\epsilon_{s,t}, \quad \epsilon_{s,t} \sim WN(0, \sigma_s^2) \\ N_t &= \psi_n(B)\epsilon_{n,t}, \quad \epsilon_{n,t} \sim WN(0, \sigma_n^2) \end{aligned}$$

としている。
X_t の時系列モデルを特徴づける $\psi(B)$ をモデル選択プロセスから特定し、幾つかの制約条件の下で $\psi(B)$ から $\psi_s(B)$ を導けば、

$$\hat{S}_t = \frac{\sigma_s^2}{\sigma^2}\frac{\psi_s(B)\psi_s(B^{-1})}{\psi(B)\psi(B^{-1})}X_t$$

の右辺のウェイトが導かれる。

4. X-13ARIMA-SEATS

■ X-12 と TRAMO-SEATS の統合

- 後述する X-13ARIMA-SEATS の AUTOMDL コマンドは、ほぼ TRAMO パートでのモデル選択の処理を踏襲している。
- 信号抽出による季節調整を実行する SEATS パートは X-13ARIMA-SEATS では **seats** コマンドとして実装されている。
- RegARIMA モデルによる処理は共通化され、回帰変数の処理や系列の延長に利用される。
- X-13ARIMA-SEATS では季節調整の方法として、X-11 フィルタと WK フィルタのいずれかを選択可能。
- **seats** コマンドにより WK フィルタを選択した場合は、RegARIMA モデルの推定結果がフィルタの導出にも利用される。

4. X-13ARIMA-SEATS

■ 新たに追加されたコマンド

- AUTOMDL
  ⇒TRAMO と同様の自動モデル選択の実行
- PICKMDL
  ⇒X-12-ARIMA の自動モデル選択プロセス
- SEATS
  ⇒SEATS の季節調整法の実行
- SPECTRUM
  ⇒ 季節性や曜日効果の事後診断を行うためにスペクトラムを出力
- FORCE
  ⇒ 原系列と季節調整系列で年間の集計値が一致するように制約をかけるオプションコマンド

4. X-13ARIMA-SEATS

■ AUTOMDL コマンド

(1) **デフォルトモデルの推定**
⇒「エアラインモデル」による回帰変数について予備的な推定

(2) **階差および季節階差の次数の特定**
⇒(1) の残差系列に対する単位根のチェック

(3) **ARMA 部分の次数の特定**
⇒ 特定された階差を適用し、ARMA 部分の次数を BIC により特定

(4) **特定されたモデルとデフォルトモデルの比較**

(5) **最終チェック**

4. X-13ARIMA-SEATS

■ (1) デフォルトモデルの推定

- ARIAM 部分を「エアラインモデル」(0 1 1)(0 1 1) に固定し、回帰部分の係数を推定する。
- 各回帰係数について t 検定で有意性を確認する。
  ⇒ サンプル数に応じた CV(critical value) を用いる。
- **outlier** コマンドで外れ値の自動選択を指定している場合はここで実行。
- 外れ値ダミーを加えても曜日効果、イースター効果、定数項が有意かチェック。
- 回帰変数が特定されると、モデルの残差に対する Ljung-Box の Q 統計量が計算される。
  （後の手順で使用）
- 回帰変数の効果を除いた系列 **linearized series** が計算される。

4. X-13ARIMA-SEATS

■ (2) 階差および季節階差の次数の特定

- **linearized series** に対して適用すべき階差を特定する。
- 一般的な単位根検定ではなく、Hannan-Rissanen 法などによって推定された ARMA モデルの AR 係数を規定の値と比較することで単位根の有無を判定。
- 単位根がある場合はそれに相当する階差を適用し、再度 ARMA モデルを推定する。
- 単位根が検出されなくなるまで手順を繰り返す。

4. X-13ARIMA-SEATS

■ (3)ARMA 部分の次数の特定

- **linearized series** に対して特定された階差操作を適用した系列について、ARMA 次数を特定する。
- 次数の上限を設定し、BIC の比較により選択する。
- 選択 S れたモデルがデフォルトモデルと異なる場合は、デフォルトモデルとの比較を行う。（手順 (4)）

4. X-13ARIMA-SEATS

■ (4) 特定されたモデルとデフォルトモデルの比較

- P_A と P_D をそれぞれ自動選択モデルとデフォルトモデルの Q 統計量の p 値とし、

$$Q_A = 1 - P_A, \quad Q_D = 1 - P_D$$

とする。

- Q_A と Q_D の値や相対的な大小関係などについての一定の規準により、自動選択モデルとデフォルトモデルのうちどちらかが選ばれる。
- 選ばれたモデルの Q が 0.975(デフォルト) を超えている場合にはモデルが不適切と判断 $\Rightarrow$ 外れ値の境界 CV を少し小さくし、(1) の外れ値の検出からやり直す。

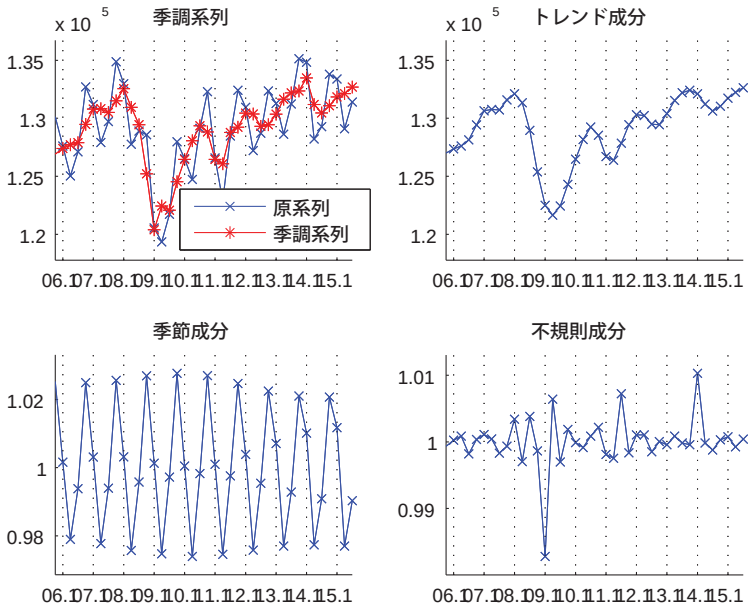
4. X-13ARIMA-SEATS

■ (5) 最終チェック

- 手順 (4) をクリアしたモデルの最終チェックが行われる。
- AR 部分の単位根の確認
  ⇒ 特性根の絶対値が 1.05 以下なら単位根と判断し、AR 次数を減らし階差を増加させる。
- MA 部分の単位根の確認
  ⇒ モデルの反転可能性を確認
- ARMA パラメータの有意性
  ⇒AR、季節 AR、MA、季節 MA のそれぞれの最大次数のパラメータが有意かどうかをチェック
  ⇒ 有意でないものがあれば、CV を小さくし手順 (1) の外れ値の検出からやり直す。

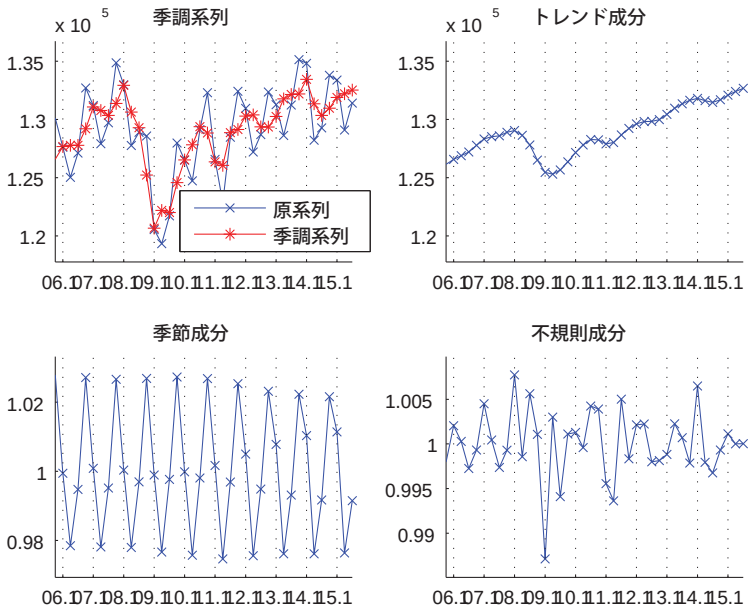
5. 調整結果の比較

■ 実質 GDP(X-11)



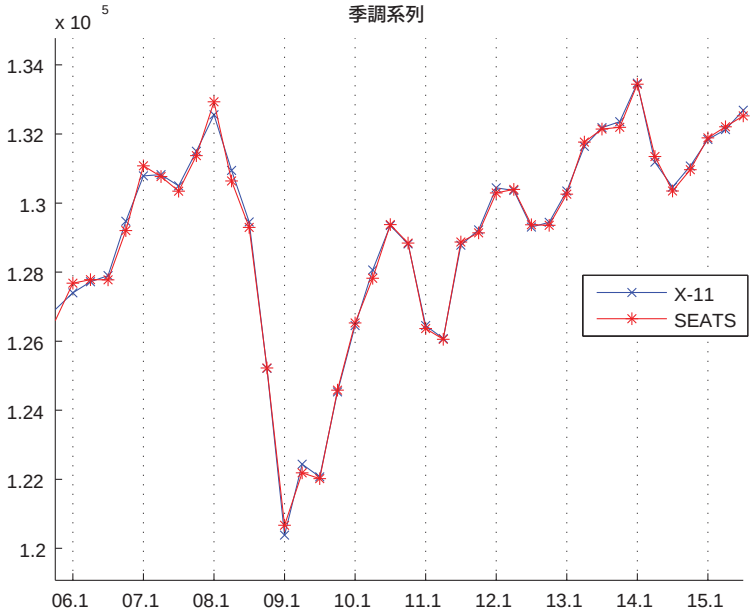
5. 調整結果の比較

■ 実質 GDP(SEATS)



5. 調整結果の比較

■ 実質 GDP(季節調整値)



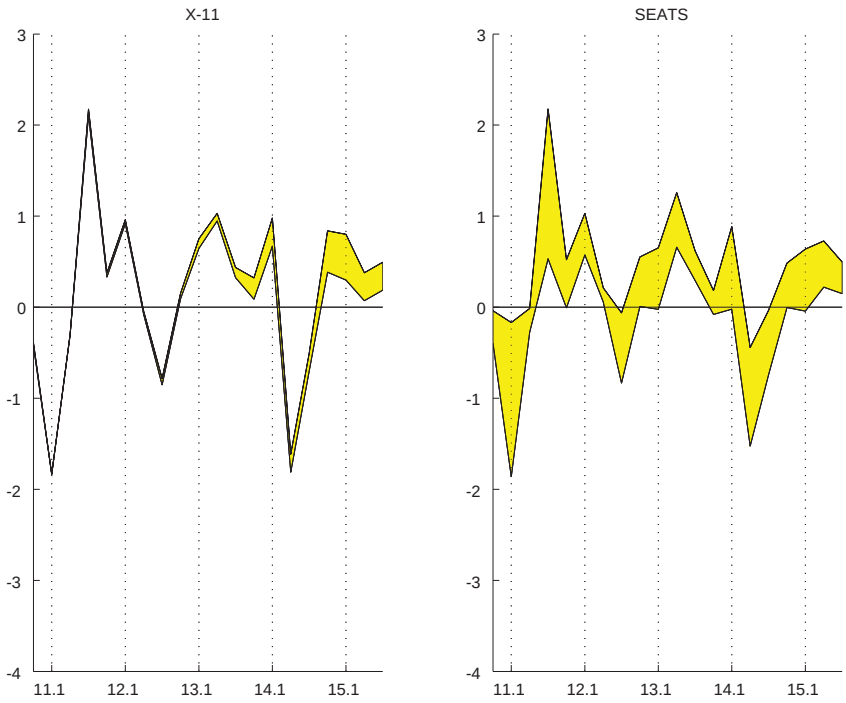
5. 調整結果の比較

■ モデル選択の影響 1

- 四半期別実質 GDP（1994 年 1-3 月～2015 年 7-9 月）
- automdl コマンドで階差次数を選択
- 階差次数を固定し、SARMA 次数をそれぞれ上限 2 として、X11 コマンドと SEATS コマンドの両ケースについて全てのモデルを推定
- それ以外のスペックはデフォルトで固定
- 各モデルによる季調値から前期比増加率を計算し、各時点での最大値、最小値を計算
- 最大値系列と最小値系列の間を黄色で着色

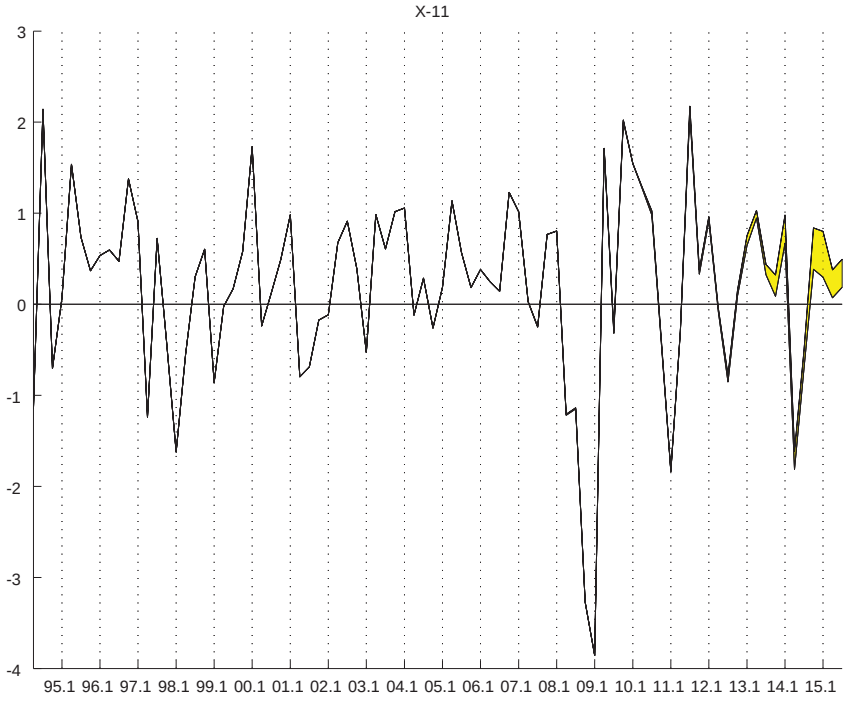
5. 調整結果の比較

■ モデル選択の影響 2



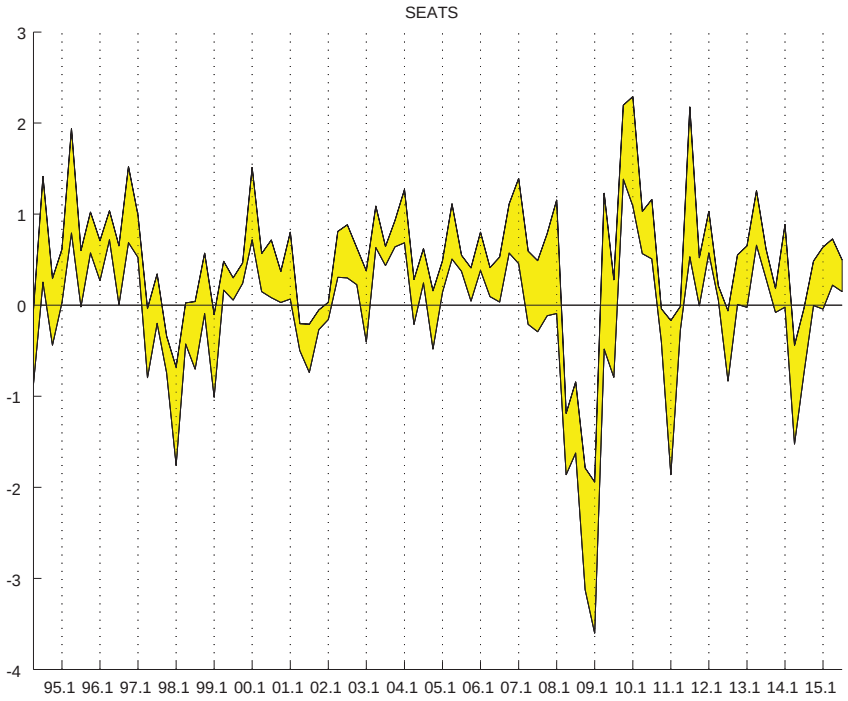
5. 調整結果の比較

■ モデル選択の影響 3



5. 調整結果の比較

■ モデル選択の影響 4



5. 調整結果の比較

■ モデル選択の影響 5

- X-11 ではモデルによる事前調整と平滑化処理が分離しているため、モデル選択結果の過去への影響は小さい。
  ⇒ ただし直近付近では最大 0.5 パーセントポイント程度の幅
- SEATS ではモデルとフィルタが連動しているため、モデルによって過去の調整値も大きく変化する。
  ⇒ 時系列的性質の変化への対応が問題
- automdl を用いた単純な季節調整では、モデル選択法が共通化されたため、X-11 と SEATS の結果差は小さい。
  ⇒ モデルの変更を伴う継続的運用においては、かなりの差異が生じる可能性
- モデル選択が不適切である場合、SEATS では季節調整が不安定になる可能性が高い。

6. まとめ

■ まとめ

- X-12-ARIMA は TRAMO-SEATS と統合される形で X-13ARIMA-SEATS に組み込まれている。
- センサス局では X-12-ARIMA のウエブ上での配布を停止しており、今後は X-13ARIMA-SEATS に情報を発信してゆくと思われる。
- X-11 フィルタと SEATS による WK フィルタの季節調整結果は、確認した範囲では非常に類似した結果を出力する。
- 一方、SEATS による処理ではモデルの変更に伴い、全期間に渡って過去の季調値に改定が生じる可能性がある。

小地域推定問題に対する ”モデルに基づくアプローチ” の新たな課題 ---海外の事例を通して---

統計数理研究所　統計思考院･データ科学研究系

廣瀬雅代

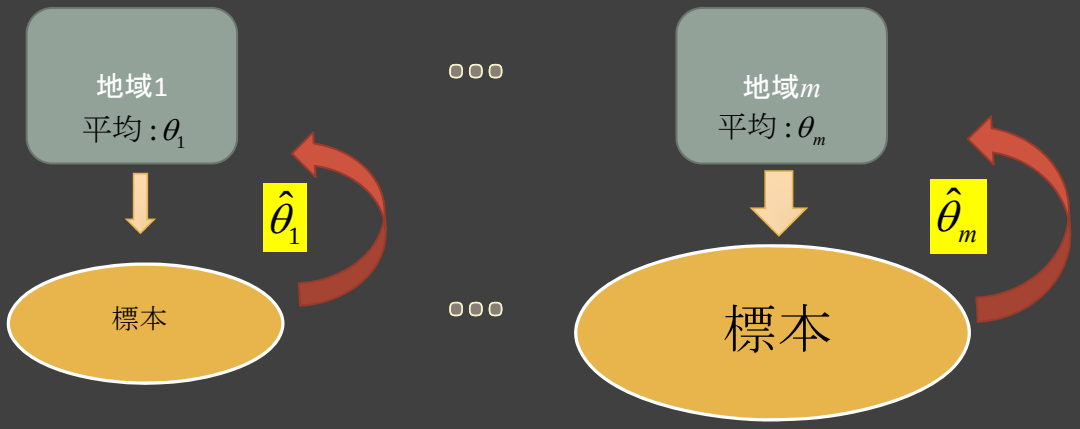
目次

1. 小地域推定問題
2. モデルに基づくアプローチと活用事例
3. モデルに基づくアプローチに対する課題と取り組み

1. 小地域推定問題

Small area estimation

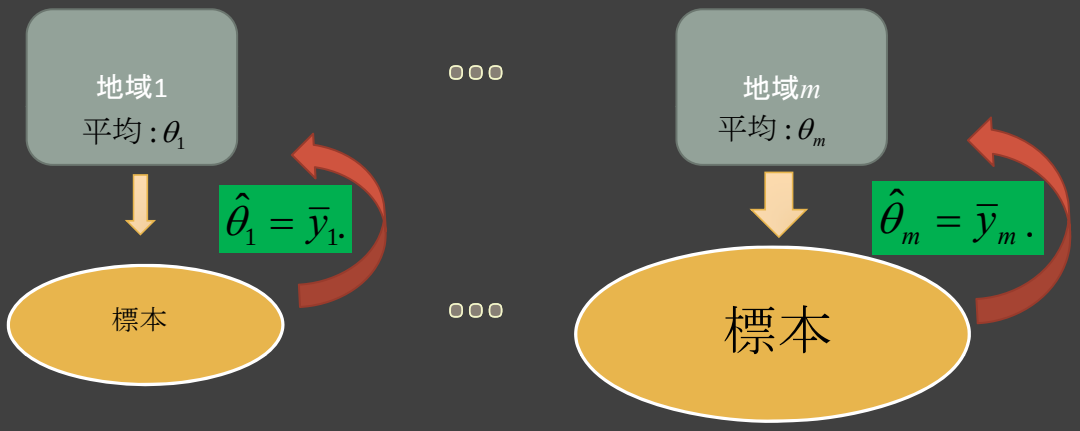
- 目的: 各地域 (州, 学区, domain) に対する特性値(平均等)の推定



Design Based approach

[Direct estimation]

- 目的: 各地域 (州, 学区, domain)に対する平均を推定



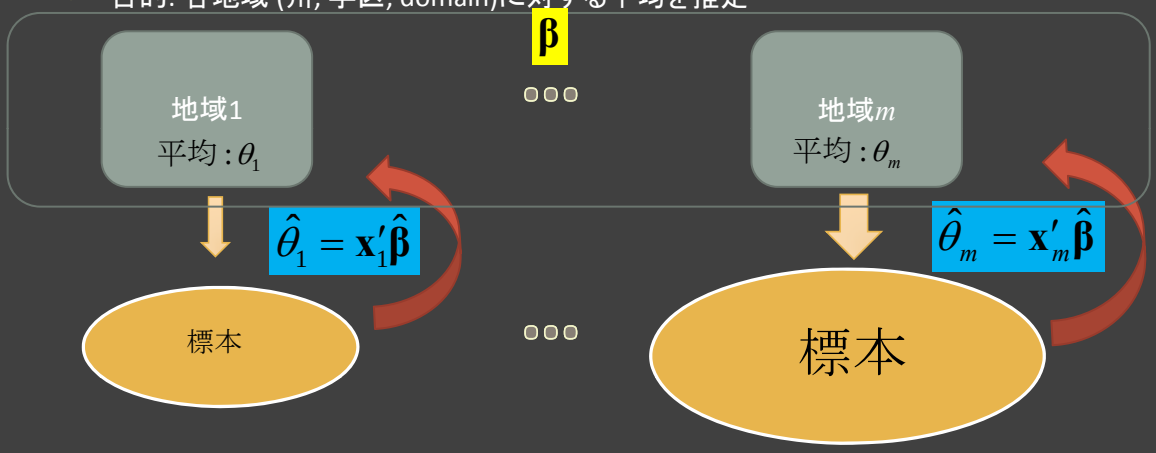
推定が不安定

モデルに基づくアプローチと活用事例

Implicit model

[Synthetic estimation/Composite estimation]

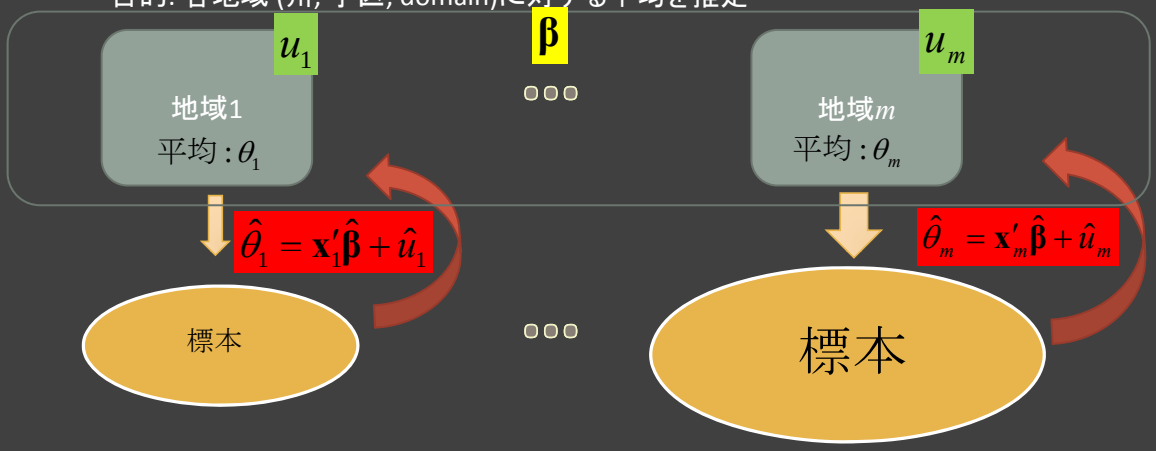
- 目的: 各地域 (州, 学区, domain)に対する平均を推定



Explicit model

[Empirical Best Linear Unbiased Predictor (EBLUP)/ Empirical Bayes (EB) /Hierarchical Bayes (HB)]

- 目的: 各地域 (州, 学区, domain)に対する平均を推定



Explicit model の有用性

Rao and Molina (2015)

1. 仮定したモデル下での最適な推定が可能
2. 各地域ごとのMSEの評価が可能
3. データからモデルを検証できる
4. 反応変数の自然な性質と複雑なデータ発生構造をモデルに組み入れることが可能
   - Spatial, time series structures

→ **EBLUP, EB, HB**

Model Based Approach

- Implicit model
- Explicit model
  - Aggregate Level model
    Unit levelのデータを使うことができないとき
  - Unit Level model
    Unit levelのデータを使うことができるとき

Aggregate level model

Direct estimates (直接推定値) と地域特有な補助情報との関係を表したモデル

-Fay Herriot model (1979)

$$g(\mathbf{y}) = \boldsymbol{\theta} + \mathbf{e},\ \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- 各地域の標本平均: $\mathbf{y}_{m\times 1} = (\bar{y}_{1.}, \cdots, \bar{y}_{m.})'$
- 各地域の平均: $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)'$
- 補助変数: $\mathbf{X}_{m\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$
- 回帰係数: $\boldsymbol{\beta} \in R^p$

  $rank(\mathbf{X}) = p < \infty,\ \sup_{i,k\geq 1} |\{\mathbf{X}\}_{ik}| < \infty$
- 地域の差異 $\mathbf{u} = (u_1, \ldots, u_m)',\ \mathbf{e} = (e_1, \ldots, e_m)'$

  それぞれ独立に $u_i \sim N(0, a),\ e_i \sim N(0, d_i),\ (d_1, \cdots, d_m$: 既知$)$

EBLUP

Unit level model

Unitごとの変数の値とunitごとの共変量との関連を表したモデル

-Nested Error Regression model (Battese et al., 1988)

$$\mathbf{y} = \boldsymbol{\theta}^* + \mathbf{e},\ \boldsymbol{\theta}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}$$

- 各Unitの値: $\mathbf{y} = (\mathbf{y}_1', \cdots, \mathbf{y}_m')'\ where\ \mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})'$
- 各domainの平均: $\boldsymbol{\theta} = (\bar{\theta}_1^*, \cdots, \bar{\theta}_m^*)'$
- 補助変数: $\mathbf{X}^*_{(N\times p)} = (\mathbf{X}^*_1, \ldots, \mathbf{X}^*_m)',\ N = \sum_{i=1}^{m} n_i$
- 回帰係数: $\boldsymbol{\beta} \in R^p$

  $rank(\mathbf{X}^*) = p < \infty,\ \sup_{i,k\geq 1} |\{\mathbf{X}^*\}_{ik}| < \infty,\ \sup_{i\geq 1} n_i < \infty$
- 地域の差異 $\mathbf{u} = (u_1, \ldots, u_m)',\ \mathbf{e} = (e_1, \ldots, e_N)',\ \mathbf{Z}^* = diag\{\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m}\}$

  それぞれ独立に $u_i \sim N(0, \sigma_u^2),\ e_{ij} \sim N(\mathbf{0}, \sigma_e^2)$

EBLUP

θ_iの予測: EBLUP

Empirical Best Linear Unbiased Predictor (EBLUP)

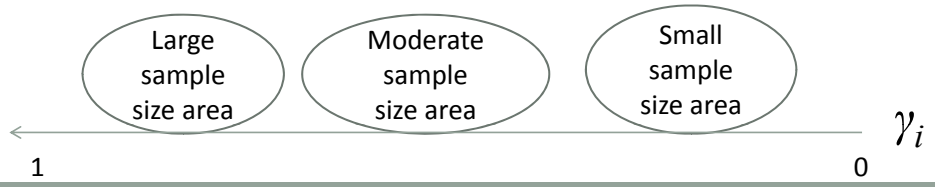
- Nested error regression model (Battesse et al., 1988; Prasad and Rao, 1990)

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i \bar{y}_i . + (1-\hat{\gamma}_i)\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},\ \hat{\mathbf{V}} = diag\{\hat{\mathbf{V}}_i\},\ \hat{\mathbf{V}}_i = \hat{\sigma}_u^2\mathbf{1}_{\mathbf{n_i}}\mathbf{1}'_{\mathbf{n_i}} + \hat{\sigma}_e^2\mathbf{I}_{\mathbf{n_i}}$$

- 縮小因子γ_i の推定量

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i}$$



事例: Per Capita Income [アメリカ]

Per Capita Income (PCI)の推定 (Fay and Herriot, 1979)

1. Design based approach [直接推定量]の活用
   - 約15000の地域で500人未満の地域
     - 約500人の地域　CV:約13%
     - 約100人の地域　CV:約30%
2. Small area placesに対して、Model based approachを用いて推定
3. Rao and Molina (2015) “1974年国勢調査局にて採用“

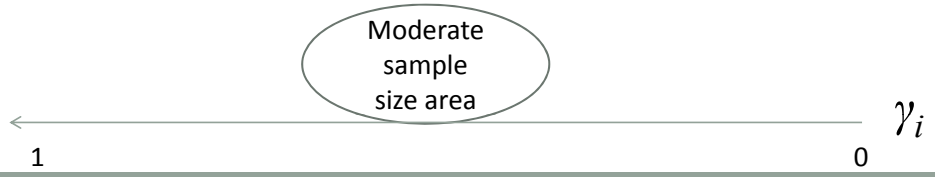
θ_iの予測: EBLUP

EBLUP

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i \bar{y}_i . + (1-\hat{\gamma}_i)\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}$$

ある正則条件下で$\mathrm{E}[(\hat{\theta}_i^{EBLUP} - \theta_i)^2] \approx \frac{1}{\gamma_i}\mathrm{E}[(\bar{y}_i - \theta_i)^2] \text{ as } m \to \infty$

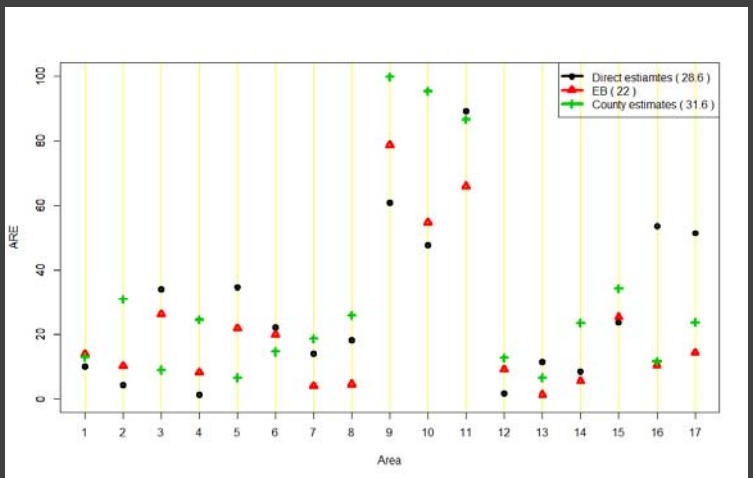
- たとえば $\gamma_i = 1/2$のとき、$\mathrm{E}[(\hat{\theta}_i^{EBLUP} - \theta_i)^2] \approx 2\mathrm{E}[(\bar{y}_i - \theta_i)^2] \text{ as } m \to \infty$



事例: Per Capita Income [アメリカ]

1973年Special complete census実施地域における1972年の値と1972年の推定値との比較

(Data source: Rao and Molina, 2015; Fay and Herriot, 1979)



$$ARE_i = \frac{|\hat{\theta}_i - T_i|}{T_i} \times 100$$

事例: Poverty counts [アメリカ]

SAIPE program in U.S.Census Bureau
(http://www.census.gov/did/www/saipe/about/index.html)

- the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program provides annual estimates of income and poverty statistics for all school districts, counties, and states.
- The main objective of this program is to provide estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions.
- In addition to these federal programs, state and local programs use the income and poverty estimates for distributing funds and managing programs.

例: Title I fund: Over $7 billion dollars of funds [Rao (2003), National research council (2000)]

- 恵まれないこどもたちのための補償教育プログラム

➡ Model based approach

モデルに基づくアプローチに対する課題と取り組み

RAO AND MOLINA (2015)

Model based approachの課題
直接推定量の活用

SAIPEにおける直接推定値

(Rao and Molina, 2015; http://www.census.gov/did/www/saipe/methods/statecounty/20102014county.html)

1. 2005年まで:人口動態調査(CPS)を基に導出

2. 2005年以後: American Community Survey (ACS)結果を基に導出
   - CPSより大規模なサンプリングが行われている

有用な直接推定値の活用

   - County estimates (約3140) : Model based approach

Model based approachの課題

Model misspecification [Model, Linking model の仮定が崩れたとき]
- Jiang et al. (2011), You and Rao (2002) etc.

元のスケールへの変換 [非線形変換を行うとバイアスが生じる]
- Slud and Maiti (2006) etc.

Benchmarkingに対する問題

$$\sum_{i=1}^{m} \omega_i \bar{y}_i \neq \sum_{i=1}^{m} \omega_i \hat{\theta}_i^{EB},$$

- You and Rao (2002) , Wang, Fuller and Qu (2008), etc.

モデルに基づくアプローチの新たな課題

Model based approachの新たな課題1

0推定値の発生 $\hat{a} = 0$

例: 1989-1992, US 5-17 years old state poverty rate (Bell, 1999)

$$\hat{\theta}_i^{\text{EBLUP}} = (1-\hat{B}_i)\bar{y}_{i\cdot} + \hat{B}_i \mathbf{x}_i'\hat{\boldsymbol{\beta}},\ \hat{B}_i = d_i/(\hat{a}+d_i)$$

$$\hat{a} = 0$$

$$\hat{B}_i = 1\ (0 < B_i < 1) \Rightarrow \hat{\theta}_i^{\text{EBLUP}} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

①非現実的 ($0<a<\infty$)

②地域の差異が反映されない

すべての地域に対してsynthetic estimatesを使用

③Parametric bootstrap 法による信頼区間利用が不可能

- Hall and Maiti (2006), Chatterjee et al.(2008)

狭義正の推定値を得る方法

Wang and Fuller (2003)

Morris and Tang (2011)

Li and Lahiri (2010)

Yoshimori and Lahiri (2014a)

Mix estimator

Mix estimator (Rubin-Bleuer and You, 2013; Molina et al., 2015)

$$\hat{a}_{Mix.LL} = \begin{cases} \hat{a}_{RE} & \text{if } \hat{a}_{RE} > 0 \\ \hat{a}_{LL} & \text{otherwise} \end{cases}$$

ある正則条件下で以下が成り立つ

$$E[\hat{a}_{Mix.LL} - a] = E[\hat{a}_{RE} - a] + o(m^{-1}), as\ m \to \infty$$

Yoshimori and Lahiri(2014a)と同様の漸近的結果

Yoshimori and Lahiri (2014a)

新たな調整項 [Under the Fay-Herriot model]

$$h_{\mathrm{YL}}(a) = \arctan[tr(\mathbf{I} - \mathbf{B})]^{1/m}$$

$$\mathbf{B} = \mathrm{diag}(B_1, \cdots, B_m),\ B_i = \frac{d_i}{a + d_i}$$

New adjusted PML estimator (AM.YL)

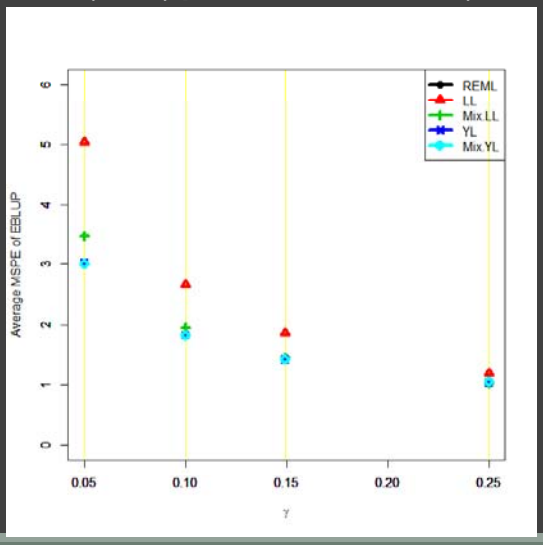
$$\hat{a}_{\mathrm{AM.YL}} = \arg\max_{0<a<\infty} h_{\mathrm{YL}}(a) L_p(a, \mathbf{y})$$

New adjusted REML estimator (AR.YL)

$$\hat{a}_{\mathrm{AR.YL}} = \arg\max_{0<a<\infty} h_{\mathrm{YL}}(a) h_{\mathrm{RE}}(a) L_p(a, \mathbf{y})$$

MIX VS YL estimator

Data Source: Rao and Molina, (2015) [Yoshimori and Lahiri (2014b)]



Model based approachの新たな課題2

計算機の利用

統計手法のソフトウェア化

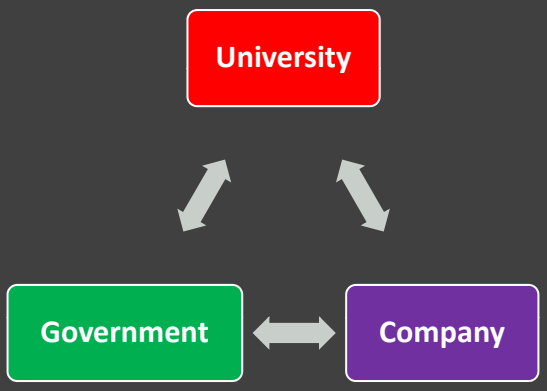
- SAS: Mukhopadhyay et al. (2011)
- R package: sae (Molina and Marhuenda, 2015) etc

計算量負荷から計算量軽減へ

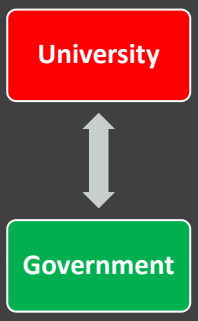
- Bootstrap methodの活用に消極的
- Yoshimori and Lahiri (2014c) Fay-Herriotモデル下での信頼区間構築

課題への取り組みと連携

アメリカ



カナダ



ヨーロッパ

The European working group on small area estimation (EWORSAE)

http://sae.wzr.pl/index.php

SAE(Small Area Estimation) conference

2001: Maryland, US

2005: Jyvaskyla, Finland

2007: Pisa, Italy

2009: Elche, Spain

2011: Trier, Germany

2013: Bangkok, Thailand

2014: Poznan, Poland

2015 : Santiago, Chile

2016: Maastricht, Netherland

Small area estimation project in Japan?

Reference (1/3)

1. Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83:** 28-36.
2. Bell, W. R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*, *52*.
3. Chatterjee, S., Lahiri, P., & Li, H. (2008). On small area prediction interval problems. *The Annals of Statistics*, **36:** 1221-1245.
4. Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74:** 269-277.
5. Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68:** 221-238.
6. Jiang, J., Nguyen, T., & Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, **106***:* 732-745.
7. Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, **101***:* 882-892.
8. Molina, I., & Marhuenda, Y. (2015). Package 'sae'.
9. Molina, I., & Morales, D. (2009). Small Area Estimation of Poverty Indicators. *Boletín de Estadística e Investigación Operativa*, **25***:* 218-225.
10. Molina, I., Rao, J. N. K., & Datta, G. S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, **41:** 1-19.

Reference (2/3)

11. Morris, C., & Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science*, **26**: 271-287.

12. Mukhopadhyay, P. K., & McDowell, A. (2011). Small area estimation for survey data analysis using SAS software. In *Proceedings of the SAS Global Forum 2011 Conference. http://support. sas. com/resources/papers/proceedings11/336-2011. pdf*.

13. National Research Council. (2000). *Small-Area Estimates of School-Age Children in Poverty:: Evaluation of Current Methodology*. In Citro, C. F. & Kalton, G. (Eds.). National Academies Press.

14. Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*,, **85:** 163-171.

15. Rao, J.N.K. (2003). *Small Area Estimation*, Wiley.

16. Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* 2nd edition, Wiley.

17. Slud, E. V., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B*, **68***:* 239-257.

18. Wang, J., & Fuller, W. A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, **98:** 716-723.

19. Wang, J., Fuller, W. A., & Qu, Y. (2008). Small area estimation under a restriction. *Survey methodology*, **34***:* 29-36.

20. Rubin-Bleuer, S. and You, Y. (2013). A Positive Variance Estimator for the Fay-Herriot Small Area Model, SRID2-12-OOIE, Statistics Canada.

Reference (3/3)

21. Yoshimori, M., & Lahiri, P. (2014a). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, **124:** 281-294.

22. Yoshimori, M., & Lahiri, P. (2014b). Supplementary material to Yoshimori, M and Lahiri, P. (2014a). Unpublished note.

23. Yoshimori, M., & Lahiri, P. (2014c). A second-order efficient empirical Bayes confidence interval. *The Annals of Statistics*, **42***:* 1-29.

24. You, Y., & Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **30**: 431-439.

正値地域データを解析するための変換モデルについて

菅澤翔之助

統計数理研究所
リスク解析戦略研究センター特任研究員

2016 年 1 月 29 日

概要

1. 基本となるモデル (Fay-Herriot モデル) について
2. 正値データのケースについて
3. unmatched sampling とリンク関数によるモデル
4. 今後の展開

1. Fay-Herriot モデル

目標: y_i を地域単位の集計データとしたとき $\theta_i = E[y_i|\theta_i]$ を推定したい.

難点: 集計数が少ないために y_i が θ_i の良い推定量になっていない.

$\Rightarrow$　モデルの力を利用して y_i よりも精度良い推定量を構成する.

そのために以下の 2 つの構造を仮定.

- $y_i|\theta_i \sim N(\theta_i, D_i)$

  y_i は真値 θ_i の周りで正規分布している.

- $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \quad v_i \sim N(0, A).$

  θ_i は地域毎の共変量 $\boldsymbol{x}_i$ と地域特有の効果 v_i によって説明される.

このモデルを Fay-Herriot モデル (Fay and Herriot, 1979) という.

1. Fay-Herriot モデル

前スライドの構造をまとめて表現すると以下のようになる.

Fay-Herriot モデル

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m, \quad v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i).$$

- D_i は既知 (実際は何らかのデータから計算する).
- $v_1, \ldots, v_m, \varepsilon_1, \ldots, \varepsilon_m$ は全て独立.
- 未知パラメータは $\boldsymbol{\beta}$ および A.
- 興味の対象: $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$.

1. Fay-Herriot モデル

- θ_i の (2 乗誤差の意味で) 最良な推定量は

$$\widetilde{\theta}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \frac{A}{A + D_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$$

  となる. この推定量は $E[(\widetilde{\theta}_i - \theta_i)^2] \leq E[(y_i - \theta_i)^2]$ を満たす.

  $\widetilde{\theta}_i$ を利用することで y_i よりも精度良い推定値を与えることができる.

- $\widetilde{\theta}_i$ は未知パラメータ $\boldsymbol{\beta}, A$ に依存しているので、実際に利用するためにはそれらを推定値 $\widehat{\boldsymbol{\beta}}, \widehat{A}$ で置き換えた $\widehat{\theta}_i$ を用いる必要がある.

  $\widehat{\boldsymbol{\beta}}, \widehat{A}$ は最尤推定法、モーメント法などの手法が提案されている.

2. 正値データのケース

現実のデータ解析では y_i が正値データのケースはよくある.

- FH モデルは $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i$ だったので各地域データ y_i は平均 $\boldsymbol{x}_i'\boldsymbol{\beta}$ の正規分布に従っていると仮定している.

  正値データは対称に分布していないケースが多い.

- 正値データに対しては y_i の代わりに $\log y_i$ に FH モデルを当てはめる.

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

  対数変換することは y_i の取りうる範囲を実数にすることと分布を対称にする２つの目的がある.

2. 正値データのケース

再掲

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

- $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$ に対して地域パラメータ $\mu_i = \exp(\theta_i)$ の推定を考える.

  この枠組みでは Slud and Maiti (2006) によって推定量およびリスク評価法が与えられている.

- 対数変換以外の変換について.

  対数変換を含むクラスの変換を考えて柔軟に推定するモデルが Sugasawa and Kubokawa (2015) で提案されている.

2. 正値データのケース

再掲

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

- このモデルは利用しやすいが実は問題がある．そもそもの目的は以下であった．

  目標: y_i を地域単位の集計データとしたとき $\theta_i = E[y_i|\theta_i]$ を推定したい．

  前スライドで興味あるパラメータを $\mu_i = \exp(\theta_i)$ と定義した．このモデルは $y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i) = \mu_i \exp(\varepsilon_i)$ と表現できるので

  $$E[y_i|\mu_i] = \mu_i \exp(D_i/2) \neq \mu_i$$

  となる．

- 単純に y_i を変換したモデルではそもそもの目的からずれたものを推定している．

3. unmatched sampling とリンク関数によるモデル

You and Rao (2002) はリンク関数を用いたモデルを提案した.

$$y_i = \theta_i + \varepsilon_i, \qquad h(\theta_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m$$

- $h(\theta_i) = \log \theta_i$ ととると対数変換モデルの代用になる.
- このモデルは $E[y_i|\theta_i] = \theta_i$ を満たす.
- You and Rao (2002) は一般のリンク関数 $h(\cdot)$ の設定でパラメータに事前分布を設定してベイズ推定を行っている.

  事前分布のパラメータを設定する必要があり、(モデル自体は有用であるが) 実用上あまり好まれない.

3. unmatched sampling とリンク関数によるモデル

本研究の目的: 頻繁に使用される対数リンク $h(\theta_i) = \log \theta_i$ のケースに限定して頻度論的な推定方法を考える.

$\Rightarrow$ 対数変換の手法に対する代替手法を提案する.

困難な点

- 以下のように y_i の周辺分布が解析的に得られない.

$$f(y_i) = \frac{1}{(2\pi D_i)^{1/2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y_i - \exp(\sqrt{A}t_i + \boldsymbol{x}_i'\boldsymbol{\beta}))^2}{2D_i} \right) u(t_i) dt_i,$$

  $u(\cdot)$ は標準正規分布の密度関数.

  最尤推定を行うのは数値積分を含んだ繰り返し計算が必要.

3. unmatched sampling とリンク関数によるモデル

再掲

$$y_i = \theta_i + \varepsilon_i, \quad \log \theta_i = \boldsymbol{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m$$
$$v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i)$$

未知パラメータ $\boldsymbol{\phi} = (\boldsymbol{\beta}', A)'$.

パラメータ推定方法: Godambe and Thompson (1989) による推定方程式を利用する.

- この推定方程式は y_i の 4 次までのモーメントが必要.
- 周辺尤度は解析的に求まらないが y_i の周辺モーメントは解析的に得られる.

3. unmatched sampling とリンク関数によるモデル

$m_i \equiv \mathrm{E}[y_i] = \exp(x_i'\boldsymbol{\beta} + A/2)$ に対して $\mu_{ki} = \mathrm{E}[(y_i - m_i)^k]$ と定義. このとき

$$\mu_{2i} = m_i^2(e^A - 1) + D_i, \quad \mu_{3i} = m_i^3(e^A - 1)^2(e^A + 2),$$
$$\mu_{4i} = m_i^4(e^A - 1)^2(e^{4A} + 2e^{3A} + 3e^{2A} - 3) + 6m_i^2 D_i(e^A - 1) + 3D_i^2.$$

さらに $u_{1i} = y_i - m_i$, $u_{2i} = (y_i - m_i)^2 - \mu_{2i}$ に対して $\boldsymbol{u}_i(y_i, \boldsymbol{\phi}) = (u_{1i}, u_{2i})'$ とし, $\boldsymbol{\Sigma}_i(\boldsymbol{\phi})$, $\boldsymbol{P}_i(\boldsymbol{\phi})$ を以下のように定義.

$$\boldsymbol{\Sigma}_i(\boldsymbol{\phi}) = \begin{pmatrix} \mu_{2i} & \mu_{3i} \\ \mu_{3i} & \mu_{4i} - \mu_{2i}^2 \end{pmatrix}, \quad \boldsymbol{P}_i(\boldsymbol{\phi})' = m_i \begin{pmatrix} \boldsymbol{x}_i & 2m_i\boldsymbol{x}_i(e^A - 1) \\ 1/2 & m_i(2e^A - 1) \end{pmatrix}.$$

このとき $\boldsymbol{\phi}$ の推定方程式は

$$\sum_{i=1}^{m} \boldsymbol{P}_i(\boldsymbol{\phi})'\boldsymbol{\Sigma}_i(\boldsymbol{\phi})^{-1}\boldsymbol{u}_i(y_i, \boldsymbol{\phi}) = \boldsymbol{0}.$$

3. unmatched sampling とリンク関数によるモデル

θ_i の推定について

- θ_i のベイス推定量は以下のようになる.

$$\widetilde{\theta}_i(y_i, \boldsymbol{\phi}) = E[\theta_i | y_i]$$
$$= \frac{\mathrm{E}_z\left[\exp\left\{\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta} - (2D_i)^{-1}(y_i - \exp(\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta}))^2\right\}\right]}{\mathrm{E}_z\left[\exp\left\{-(2D_i)^{-1}(y_i - \exp(\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta}))^2\right\}\right]},$$

  $E_z[\cdot]$ は $z \sim N(0,1)$ に対する期待値を表す.

- 推定量を代入することで最終的に $\widehat{\theta}_i = \widetilde{\theta}_i(y_i, \widehat{\boldsymbol{\phi}})$ を得る.(積分の部分は解析的に計算できないので数値積分で評価する必要がある.)

3. unmatched sampling とリンク関数によるモデル

$\widehat{\theta}_i$ のリスク評価

$\widehat{\theta}_i$ のリスク評価のために $\widehat{\theta}_i$ の MSE を考える.

$$\mathrm{MSE}_i = E[(\widehat{\theta}_i - \theta_i)^2]$$

- 一般に MSE_i は未知パラメータ $\boldsymbol{\phi}$ に依存するので、MSE_i の精度良い推定量を用いる.

  具体的には MSE の推定量を $\widehat{\mathrm{MSE}}_i$ が $E\left[\widehat{\mathrm{MSE}}_i\right] = \mathrm{MSE}_i + o(m^{-1})$ を満たすように構成する.

- 推定方程式で定義した $\widehat{\boldsymbol{\phi}}$ の漸近分散、漸近バイアスを評価して解析的に求めることができる. またパラメトリックブートストラップを用いて構成することもできる.

3. unmatched sampling とリンク関数によるモデル

数値実験

対数リンクが真のとき、対数変換モデルはどのくらい機能するのか.

- データ生成過程

$$
\begin{aligned}
y_i =& \theta_i + \varepsilon_i, \quad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 30 \\
& v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i)
\end{aligned}
$$

  $\beta_0 = 0, \beta_1 = 0.6,\ x_i \sim N(0, 5),\ D_i \sim U(5, 15),\ A = 0.5, 1.$

- θ_i の予測量を対数リンク、対数変換、FH のそれぞれから計算し、真の θ_i との MSE および bias を 1,000 回の繰り返から計算.

  また direct estimator y_i の MSE および bias も同様に計算.

3. unmatched sampling とリンク関数によるモデル

数値実験 (結果)

		対数リンク	FH	対数変換	y_i
$A = 1$	MSE	2.62	3.01	3.02	3.05
	bias	0.145	0.325	0.112	0.332
$A = 0.5$	MSE	2.63	2.76	2.78	2.86
	bias	-0.477	0.178	-0.046	0.192

3. unmatched sampling とリンク関数によるモデル

実データへの適用

- y_i: 2014 年の都道府県ごとの家計調査 (教育費), $\quad i = 1, \ldots, 47$
- D_i は 2006 年から 2013 年のデータから計算.
- 共変量 x_i として 2011 年の大規模家計調査の結果を利用する.

対数リンクモデルを当てはめる.

$$y_i = \theta_i + \varepsilon_i, \qquad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 47$$

3. unmatched sampling とリンク関数によるモデル

実データへの適用

$$y_i = \theta_i + \varepsilon_i, \quad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 47$$

推定値: $\widehat{\beta} = 0.984,\ \widehat{\beta} = 0.843,\ \widehat{A} = 3.16$

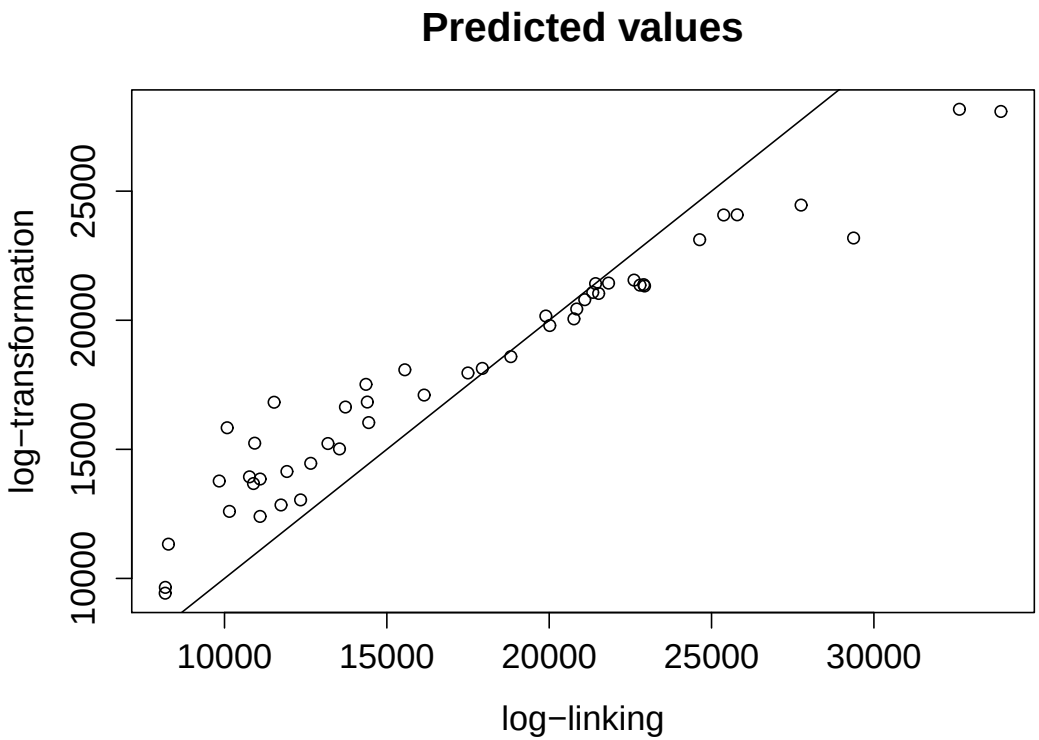


Figure: 対数リンクモデルと対数変換モデルの予測値

4. 今後の展開

考察

- unmatched sampling と対数リンクを用いたモデルは正値データに対して有用.
- 現実には割合などの有界連続値をとるデータもある. また実数値データでも非対称に分布しているケースもあるかもしれない.

  logistic リンク? $\mathbb{R}$ から $\mathbb{R}$ のリンク?
- そもそも対数リンクなど決め打ちしたリンクが適当とは限らない.

  リンクもデータから推定できると良い.

4. 今後の展開

ノンパラメトリックリンクを用いた unmatched-sampling モデル

$$y_i = \theta_i + \varepsilon_i, \qquad \theta_i = L(\boldsymbol{x}_i^t \boldsymbol{\beta} + v_i), \quad i = 1, \ldots, m.$$
$$v_i \sim N(0, 1), \quad \varepsilon_i \sim N(0, D_i)$$

リンク $L(\cdot)$ は以下のように P-spline を用いて表現する.

$$L(u) = \gamma_0 + \gamma_1 u + \ldots + \gamma_q u^q + \sum_{k=1}^{K} \gamma_{q+k} (u - t_k)_+^q.$$

- 関連する統計モデル: single index モデル (ただし既存のものは random effect を含まないもののみ)
- パラメータを頻度論的に推定するのは難しいが, 事前分布を設定してベイズ推定するのは比較的容易.(事後分布からのサンプリングは簡単に実行可能.)

  客観性を保つために $\boldsymbol{\beta}, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q+K})'$ に uniform prior を入れる.

4. 今後の展開

再掲

$$y_i = \theta_i + \varepsilon_i, \qquad \theta_i = L(\boldsymbol{x}_i^t \boldsymbol{\beta} + v_i), \quad i = 1, \ldots, m.$$

$$L(u) = \gamma_0 + \gamma_1 u + \ldots + \gamma_q u^q + \sum_{k=1}^{K} \gamma_{q+k} (u - t_k)_+^q.$$

- uniform prior を設定した場合, 事後分布は必ずしも proper にならない.

  適当な条件のもと事後分布は proper になることを示した.

- どんな連続値データに対しても利用可能.

  数値実験では対数リンクが misspecify されたケースで対数リンクに対する優越性なども観察された.

まとめ

- データを変換するモデルの代用として unmatched sampling とリンク関数によるモデルを導入した.
- 対数変換のケースに焦点を当て, 頻度論の枠組みでモデルパラメータの推定や小地域パラメータの推定, およびそのリスク評価について提案した.
- 今後の主展開: ノンパラメトリックリンクを用いたモデル

参考文献

- Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.
- Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference*. **22**, 137-152.
- Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of Royal Statistical Society: Series B*, **68**, 239-257.
- Sugasawa, S. and Kubokawa, T. (2015). Parametric transformed Fay-Herriot model for small area estimation. *Journal of Multivariate Analysis*, **139**, 295-311.
- You, Y. and Rao, J. N. K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, **30**, 3-15.

空間重み付き経験ベイズ推定と死亡データへの応用

川久保友超

千葉大学・法政経学部

2016 年 1 月 29 日

科研コンファレンス「経済統計・政府統計の理論と応用」

導入

要約

- 地理情報を組み込んだ小地域推定のモデルを提案.
- 提案モデルは，カウントデータや二値データなどの離散データにも適用できる.
- 死亡データへの応用を行い，標準化死亡比（SMR）と呼ばれるリスク指標を小地域推定する.
- 本発表の内容は，以下の working paper にもとづいている.

  Sugasawa, S., Kawakubo, Y. and Ogasawara, K.
  Geographically weighted empirical Bayes estimation via natural exponential family.
  arXiv preprint, arXiv:1508.01641.

導入

小地域推定

- 地域 i の所得の平均 μ_i を知りたいとき，サンプルサイズが小さい地域では，集計データ（標本平均）y_i は推定誤差が大きく信頼できない.
- μ_i を小地域母数，y_i をその direct estimator という.
- μ_i に周辺地域の情報を入れた事前分布を仮定すると，μ_i の経験ベイズ推定量は安定した推定量となる（borrowing strength）

  → model based estimator
- 線形混合モデル（Linear Mixed Model, LMM）と呼ばれるモデルのクラスが最も広く使われている.

導入

- カウントデータや二値データなどの離散データに対しては LMM は適切でないため，より広いクラスの一般化線形混合モデル（GLMM）がしばしば用いられる.
- しかしながら，GLMM は推定に数値積分が必要で，実行上煩雑.
- GLMM に代わり，小地域母数に共役事前分布を仮定したモデルを用いる．このモデルは，解析的にベイズ推定量が導出できる等のメリットがある.

既存のモデル

- Ghosh and Maiti (2004, *Biometrika*) によって次のモデルが提案された.

  $y_1, \ldots, y_m$ は独立. $y_i|\theta_i$ および θ_i の分布を以下のように設定.

$$y_i|\theta_i \sim f(y_i|\theta_i) = \exp[n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)],$$
$$\theta_i \sim \pi(\theta_i|\nu, m_i) = \exp[\nu(m_i\theta_i - \psi(\theta_i))]C(\nu, m_i),$$

  n_i: 既知

  $m_i = \psi'(\boldsymbol{x}_i^t\boldsymbol{\beta})$　　ハイパーパラメータ: $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \nu)^t$

  興味の対象: $\mu_i = E[y_i|\theta_i]$.

- 分散構造として $\mathrm{Var}(y_i|\theta_i) = n_i^{-1}Q(\mu_i) = n_i^{-1}(v_0 + v_1\mu_i + v_2\mu_i^2)$ を仮定.

  これらは正規分布、二項分布、ポアソン分布を含むため実用上十分広いクラス.

既存のモデル

- θ_i には共役事前分布を入れているので周辺尤度および事後分布は解析的に求まり，μ_i のベイズ推定量として以下を得る.

$$\widetilde{\mu}_i = \widetilde{\mu}_i(y_i, \boldsymbol{\phi}) = \frac{n_i y_i + \nu m_i}{n_i + \nu}.$$

この推定量は未知のパラメータ $\boldsymbol{\phi}$ に依存するので infeasible

- $\boldsymbol{\phi}$ は周辺尤度から以下のように推定できる.

$$\widehat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{i=1}^{m} \left[\log C(\nu, m_i) - \log C(n_i + \nu, \widetilde{\mu}_i(y_i, \boldsymbol{\phi})) \right].$$

この $\widehat{\boldsymbol{\phi}}$ をベイズ推定量 $\widetilde{\mu}_i$ に代入することで経験ベイズ推定量 $\widehat{\mu}_i = \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}})$ を得る. → model based estimator

動機

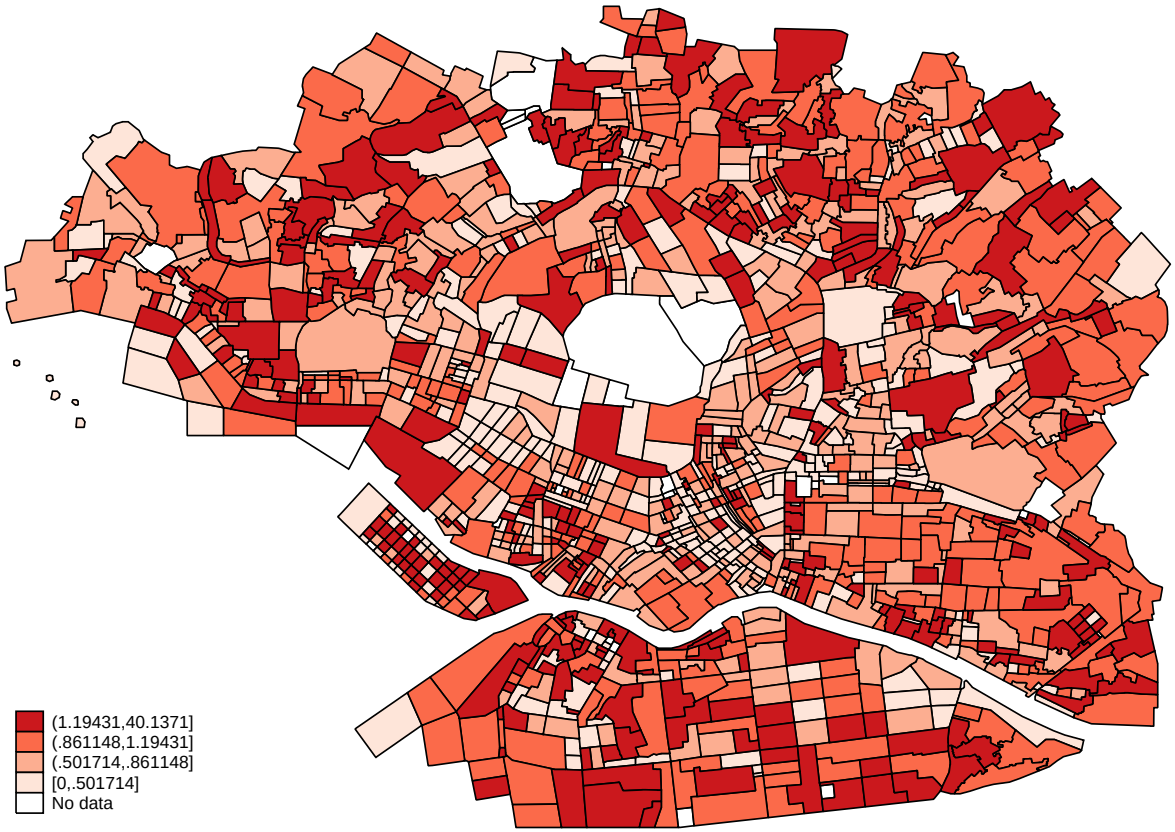


Figure : 1930 年東京市における男性の SMR の分布

問題提起

- 既存のモデルではハイパーパラメータが各地域で共通であることを仮定していた.

  ハイパーパラメータは各地域で異なっている (空間非定常) と考えるのが自然.

- 各地域の地理的関係性が情報として得られている場合，それもモデルに組み込むことができた方が精度良い推定ができそう.

空間重み付き経験ベイズ推定

- 既存モデルにおいて $\boldsymbol{\phi} = \boldsymbol{\phi}_i$ としたモデルを提案する.

  $\boldsymbol{\phi}_i,\ i = 1, \ldots, m$ を以下のように推定

$$\widehat{\boldsymbol{\phi}}_i = \operatorname*{argmax}_{\boldsymbol{\phi}} \sum_{k=1}^{m} w_{ik} \left[\log C(\nu, m_k) - \log C(n_k + \nu, \widetilde{\mu}_k(y_k, \boldsymbol{\phi})) \right],$$

$$m_k = \psi'(\boldsymbol{x}_k^t \boldsymbol{\beta}), \quad \widetilde{\mu}_k(y_k, \boldsymbol{\phi}) = (n_k y_k + \nu m_k)/(n_i + \nu).$$

  w_{ik} は２つの地域 i, k 間のウェイトで、d_{ik} を地域間の距離，b をバンド幅としたときに，カーネル関数 $K(x)$ を用いて $w_{ik} = K(d_{ik}/b)$ と定義する.

- カーネル関数として対称かつ有界サポートをもつカーネルを用いる．今回は以下の 4 次カーネルを用いる.

$$K(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & (0 \leq x \leq 1), \\ 0, & (x > 1). \end{cases}$$

空間重み付き経験ベイズ推定

- 地域毎に推定された $\widehat{\phi}_i$ を用いて空間重み付き経験ベイズ（GWEB）推定量を

$$\widehat{\mu}_i^{\mathrm{GWEB}} = \frac{n_i y_i + \widehat{\nu}_i \widehat{m}_i}{n_i + \widehat{\nu}_i},$$

  と定義.

- バンド幅 b は以下のような Cross Validation による基準を用いて選択する.

$$\mathrm{CV}(b) = \sum_{i=1}^{m} \left\{ y_i - \widehat{\mu}_{(-i)}^{\mathrm{GWEB}}(b) \right\}^2,$$

  ただし $\widehat{\mu}_{(-i)}^{\mathrm{GWEB}}(b)$ は、バンド幅 b のもとで $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m$ から求めた $\widehat{\phi}_i$ を用いた μ_i の GWEB 推定量.

推定手順

推定ステップ

1. 各地域の距離 d_{ik} を計算する.
2. CV 基準を最小化するような最適なバンド幅 b^* を求める.
3. バンド幅 b^* のもとで各地域間のウェイト w_{ik} を計算し、各 i に対して $\widehat{\boldsymbol{\phi}}_i$ を求める.
4. 推定量 $\widehat{\boldsymbol{\phi}}_i$ をベイズ推定量に代入して $\widehat{\mu}_i^{\mathrm{GWEB}}$ を得る.

MSE

- $\widehat{\mu}_i^{\mathrm{GWEB}}$ のリスク評価のために MSE の推定量を求める.
- MSE を以下のように分解.

$$\begin{aligned}
\mathrm{MSE}_i &= E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \mu_i)^2\right] \\
&= E\left[(\widetilde{\mu}_i - \mu_i)^2\right] + E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \widetilde{\mu}_i)^2\right] \\
&= \frac{\nu_i Q(m_i)}{(n_i + \nu_i)(\nu_i - v_2)} + E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \widetilde{\mu}_i)^2\right].
\end{aligned}$$

- 第 1 項は $O(1)$, 第 2 項は $O(n^{-1})$ $(n = mb)$.
- parametric bootstrap を用いて MSE の 2 次漸近不偏推定量を求めることができる.

死亡データへの応用

データセット

- 1930 年東京市の死亡データを解析する.
- z_i: 地域 i の男性の死亡数
- N_i: 地域 i に住んでいる男性の総数
- 地域数は $m = 1372$ $(i = 1, \ldots, m)$.
- 男性の総死亡数は $L = \sum_{i=1}^{m} z_i = 13656$.
- 地域 i の男性の期待死亡数は, $n_i = L \times (N_i / \sum_{j=1}^{m} N_j)$

標準化死亡比（Standardized Mortality Ratio, SMR）

- SMR はある地域における死亡の潜在的な危険性をあらわす指標で, 「実際の死亡数と期待死亡数の比」で定義される.
- 地域 i における SMR の direct estimator は, $y_i = z_i / n_i$.
- しかしながら, 小地域では y_i は信頼できないので, SMR の model based estimator（GWEB 推定量）を求めたい.

死亡データへの応用

空間ポアソン・ガンマ混合モデル

- $z_1, \ldots, z_m$ は独立に以下の分布に従っていると仮定

$$z_i|\mu_i \sim \mathrm{Po}(n_i\mu_i), \quad \mu_i \sim \mathrm{Ga}(\nu_i m_i, \nu_i)$$

- “真の”SMR は μ_i，その direct estimator は $y_i = z_i/n_i$.
- 共変量は $\boldsymbol{x}_i = (x_{0i}, x_{1i})^t$, $x_{0i} = 1$, x_{1i} は地域 i の女性の死亡数,
  $E(\mu_i) = m_i = \exp(\boldsymbol{x}_i^t\boldsymbol{\beta}_i)$.
- 地理情報として隣接行列が得られているので，隣接行列によって定義されるグラフの最短距離として地域間の距離 d_{ik} を計算．これにもとづき
  $\boldsymbol{\phi}_i = (\beta_{0i}, \beta_{1i}, \nu_i)^t$ を推定し，$\widehat{\mu}_i^{\mathrm{GWEB}}$ を導出.

$$\widehat{\mu}_i^{\mathrm{GWEB}} = \frac{n_i}{n_i + \widehat{\nu}_i} y_i + \frac{\widehat{\nu}_i}{n_i + \widehat{\nu}_i} \exp(\boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}_i)$$

- 比較のため，既存のポアソン・ガンマ混合モデル（$\boldsymbol{\phi}_i = \boldsymbol{\phi}$）も適用.

推定結果

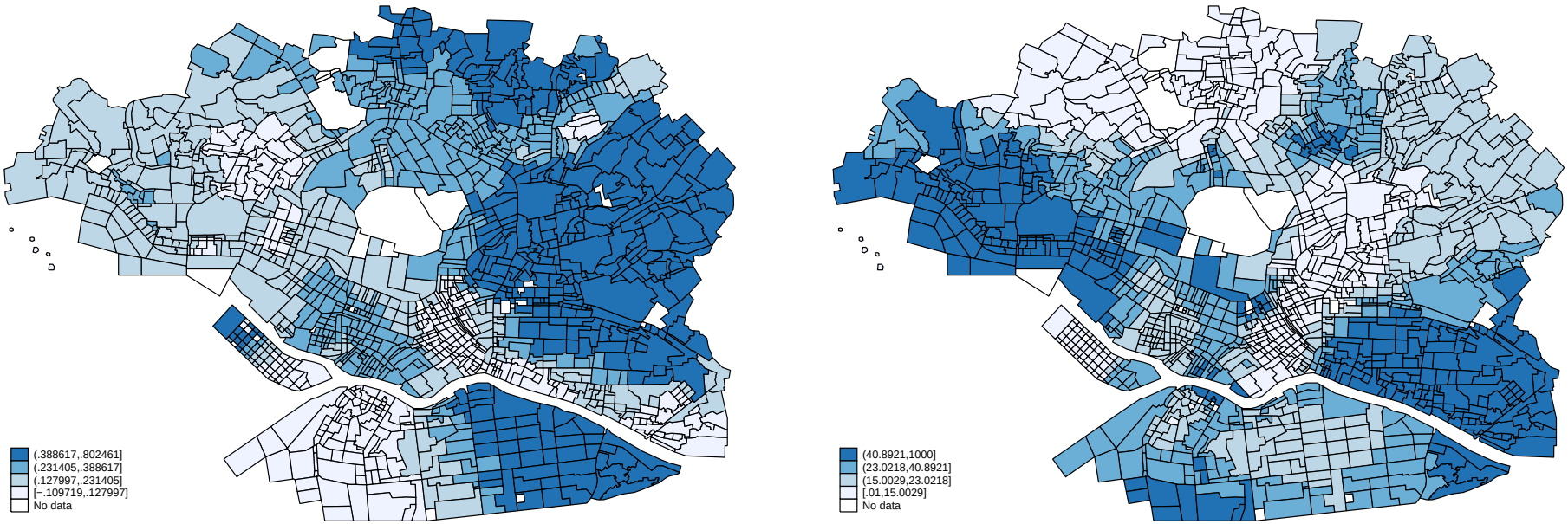


Figure : 推定されたハイパーパラメータの空間分布. 左 : β_{1i}, 右 : ν_i.

推定結果

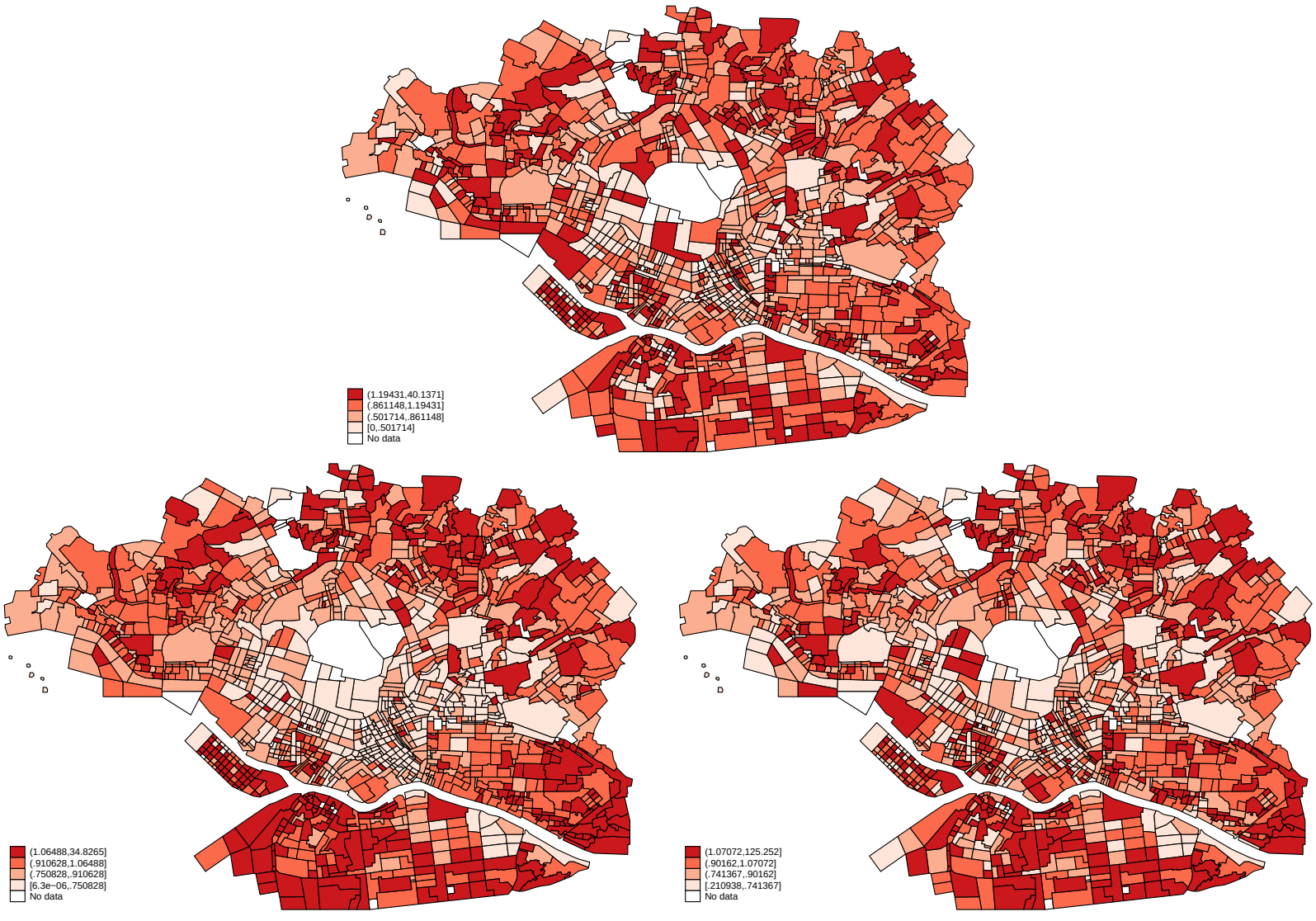


Figure : y_i（上），GWEB（左下），EB (右下)

MSE の比較

既存のポアソン・ガンマ混合モデルと比較するため，提案モデルと既存モデルの MSE の推定値を計算し，direct estimator y_i の MSE との比をそれぞれ計算.

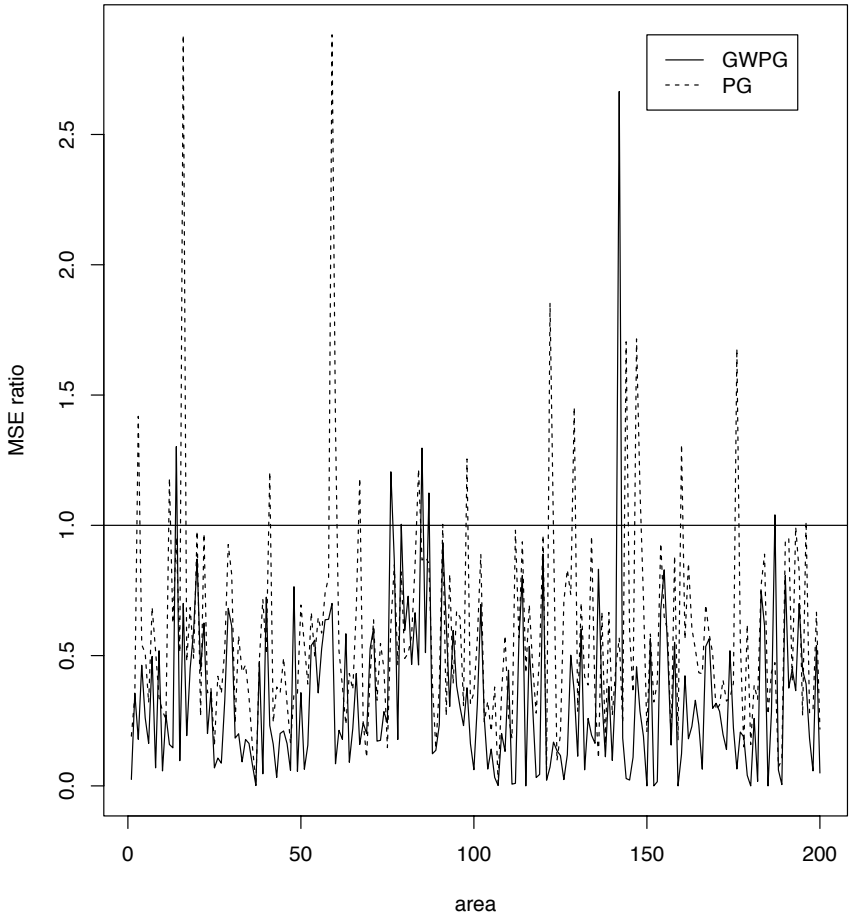


Figure : 200 地域における MSE 比.

まとめ

- 空間非定常性を考慮した小地域推定モデルを考え，そこから得られる model based estimator として GWEB 推定量を提案した.
- 提案モデルは離散データへの応用が可能.
- 例として，空間ポアソン・ガンマ混合モデルを用いて，SMR の小地域推定を行った.