地域メッシュ別標準化女性子ども比による標本抽出、推定、誤差に関する分析

（公財）統計情報研究開発センター

はじめに

- 本資料は国勢調査の地域メッシュ統計データを利用して、標本抽出、推定、誤差について分析した結果をまとめたものである。
- 分析の目的は、集落抽出における母平均の推定及びその分散の漸近性について考察することである。
- 抽出単位（要素）を標準地域メッシュ別データ、集落を市区町村境界とし、単純無作為抽出における推定との比較を行う。
- 統計項目としては標準化女性子ども比のデータを利用する。この指標の統計的な分布は、平均を中心として概ね左右対称の形状を示しており、基礎的な分析に適しているためである。
- 本資料は、標本抽出、推定、誤差評価の基本的な考え方と分析について現時点の結果をまとめたものである。8分の1地域メッシュ別のデータを用いた分析など最終的な成果は、後日報告予定である。

内容

- 地域メッシュ統計について
- 出生力の指標と標準化女性子ども比
- 標準化女性子ども比の計算
- 標本調査の基礎と単純無作為抽出の概要
- 集落抽出法の概要
- 地域メッシュ統計を利用した集落抽出における平均の推定とその分散の漸近性
- まとめ

地域メッシュ統計について

センサスの集計地域区分

- 全国を対象した集計地域区分は、大きく以下の3つに分けられる。

	令和2年国勢調査	令和3年経済センサス－活動調査
行政地域による地域区分	全国、都道府県別、市区町村別	
行政地域を細分化した地域区分	基本単位区別 町丁・字等別	町丁・大字別
緯度経度で区切られた地域区分	地域メッシュ統計 最小：4分の1地域メッシュ	地域メッシュ統計（未定） 最小：2分の1地域メッシュ

〇基本単位区別集計は最小の地域区分別であり、詳細な地域単位別の統計結果を確認できる。しかし、集計項目は「男女別人口及び世帯数」となっており限定的である。
〇町丁・字等別集計は、地域の単位は基本単位区よりも大きいが、世帯、住居、産業、職業、従業地・通学地、移動など、様々な項目について確認できる。
〇経済センサスー活動調査の町丁・大字別集計の集計項目は、産業(中分類)別・経営組織別・従業者規模別事業所数及び従業者数となっている。

地域メッシュ統計とは

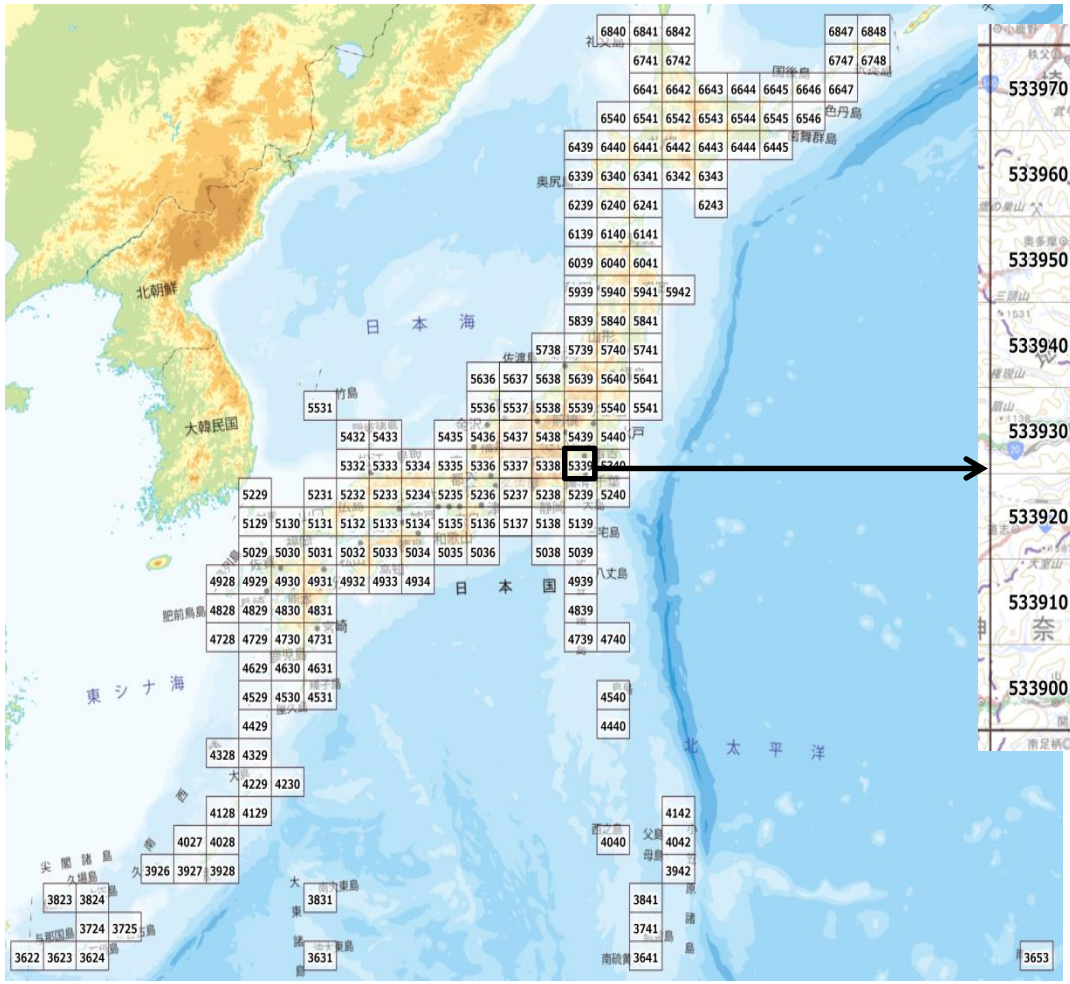
- 地域メッシュ統計
  - 地域メッシュ統計とは、緯度・経度に基づき地域を隙間なく網の目（メッシュ）の区域に分けて、それぞれの区域に関する統計データを編成したものである。

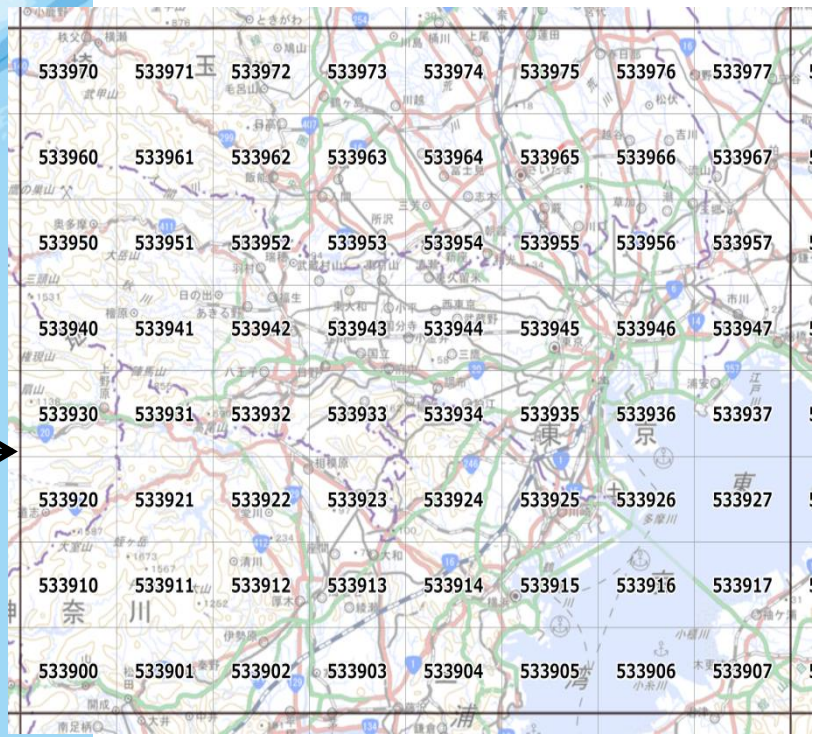
- 地域メッシュ統計の利点
  - 時系列比較がしやすい
    - 緯度・経度に基づいて区域が分けられているため、その位置や区画が固定されていることから、市区町村などの行政区画の境界変更や地形、地物の変化による調査区の設定の変更などの影響を受けない。
  - 地域メッシュ相互間の事象の計量的比較がしやすい
    - 日本は中緯度に位置しており、メッシュ区画の大きさは全国ほぼ同じ大きさである。

地域メッシュの区分方法

区画の種類	区分方法	緯度の間隔	経度の間隔	一辺の長さ	地図との関係	編成範囲
第1次地域区画	全国の地域を偶数緯度及びその間隔（120分）を3等分した緯度における緯線並びに1度ごとの経線とによって分割してできる区域	40分	1度	約80km	20万分の1地勢図の1図葉の区画	全国
第2次地域区画（統合地域メッシュ）	第1次地域区画を緯線方向及び経線方向に8等分してできる区域	5分	7分30秒	約10km	2万5千分の1地勢図の1図葉の区画	全国
基準地域メッシュ（第3次地域区画）	第2次地域区画を緯線方向及び経線方向に10等分してできる区域	30秒	45秒	約1km		全国
2分の1地域メッシュ（分割地域メッシュ）	基準地域メッシュ（第3次地域区画）を緯線方向，経線方向に2等分してできる区域	15秒	22.5秒	約500m		全国
4分の1地域メッシュ（分割地域メッシュ）	2分の1地域メッシュを緯線方向，経線方向に2等分してできる区域	7.5秒	11.25秒	約250m		全国
8分の1地域メッシュ（分割地域メッシュ）	4分の1地域メッシュを緯線方向，経線方向に2等分してできる区域	3.75秒	5.625秒	約125m		全国

地域メッシュの区画と区分の方法





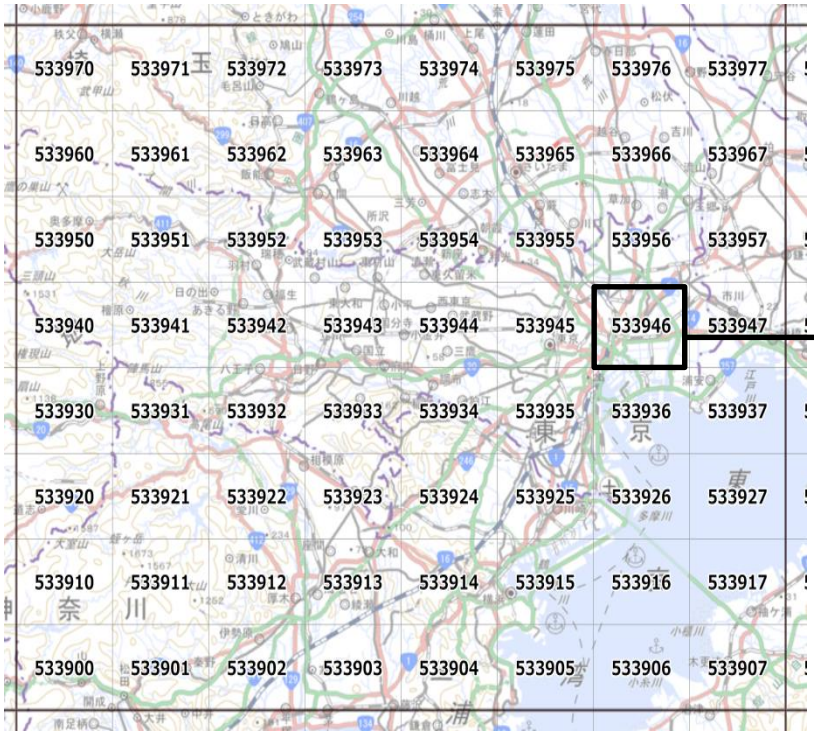
第2次地域区画（約10km）

第1次地域区画を縦横8等分

第1次地域区画（約80km）

データ名	出所
地理院タイル	国土地理院『電子地形図（タイル）』

地域メッシュの区画と区分の方法



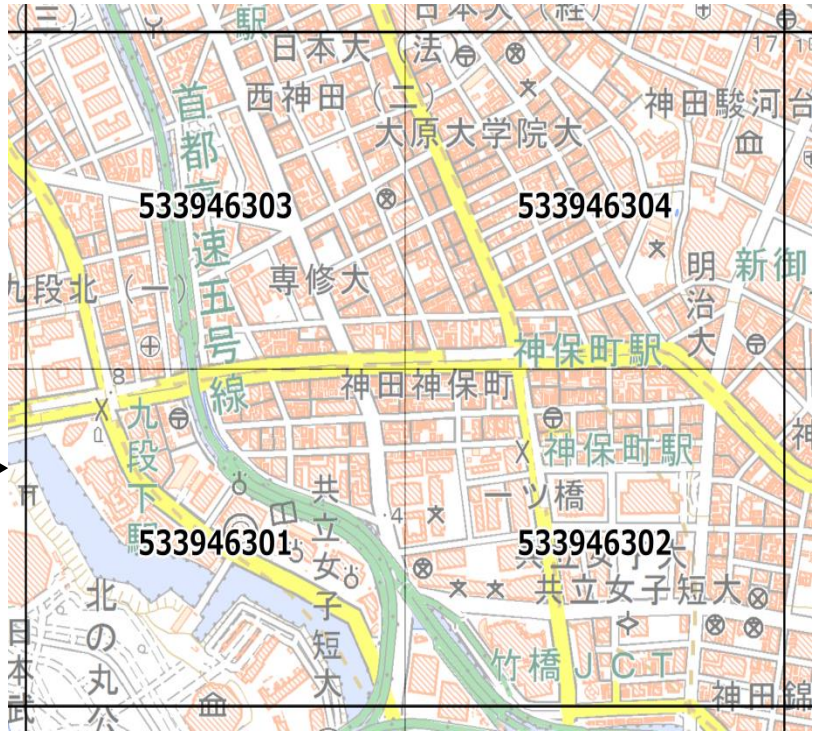
第2次地域区画（約10km）

基準地域メッシュ（第3次地域区画、約1km）
第2次地域区画を縦横10等分

データ名	出所
地理院タイル	国土地理院『電子地形図（タイル）』

地域メッシュの区画と区分の方法

基準地域メッシュ（第3次地域区画、約1km）

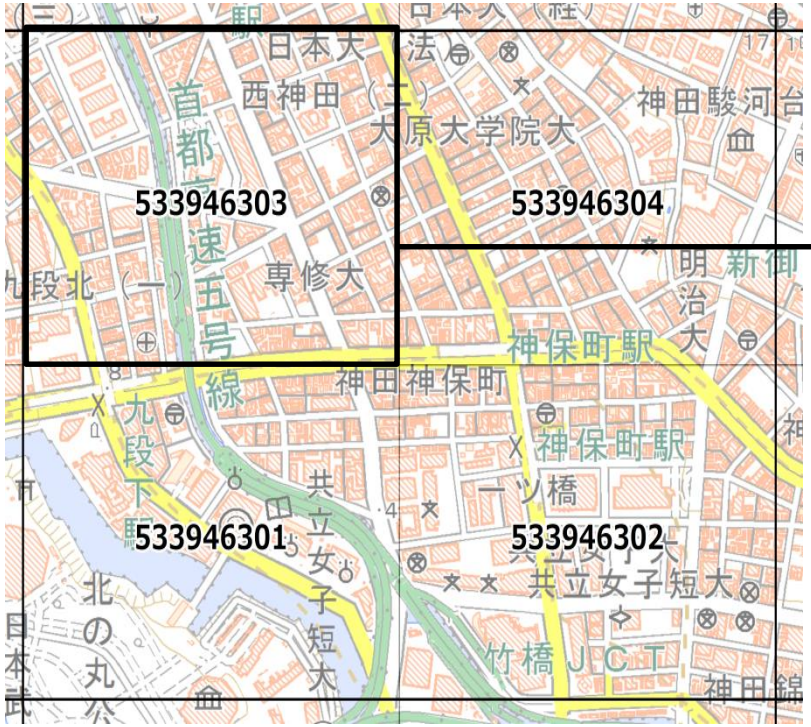


2分の1地域メッシュ（約500m）

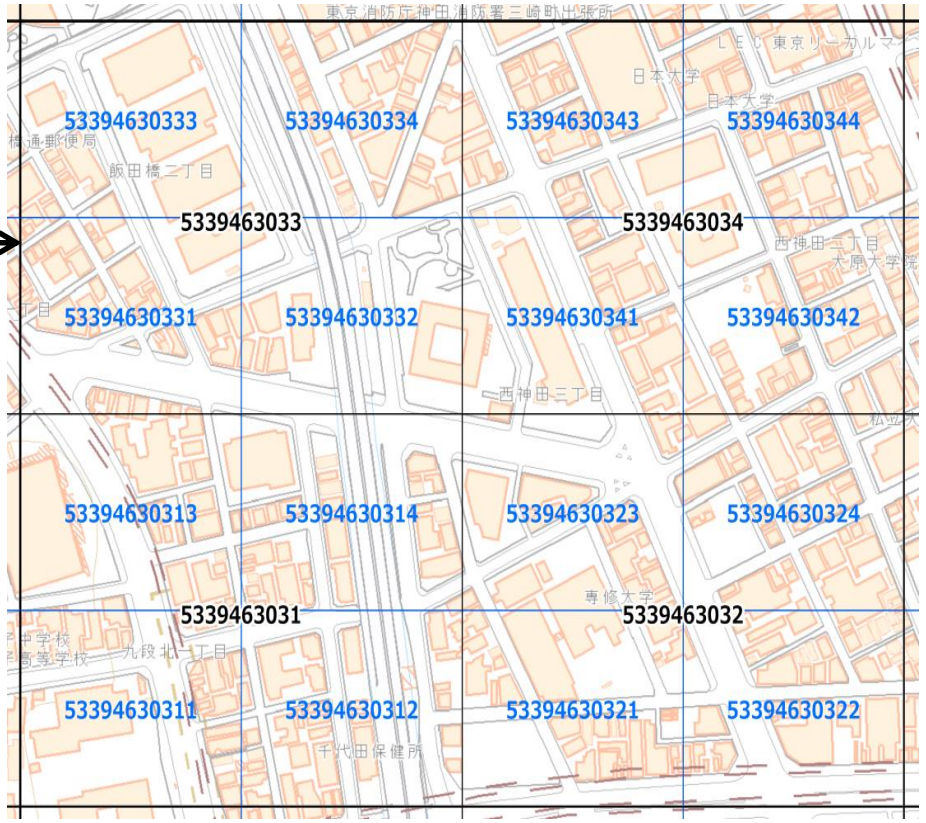
基準地域メッシュを縦横2等分

データ名	出所
地理院タイル	国土地理院『電子地形図（タイル）』

地域メッシュの区画と区分の方法



2分の1地域メッシュ（約500m）

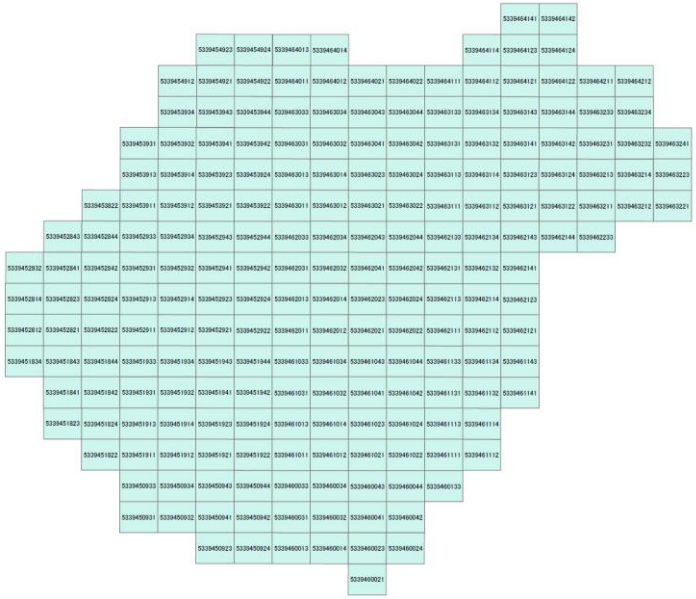


4分の1地域メッシュ（約250m）
２分の1地域メッシュを縦横2等分

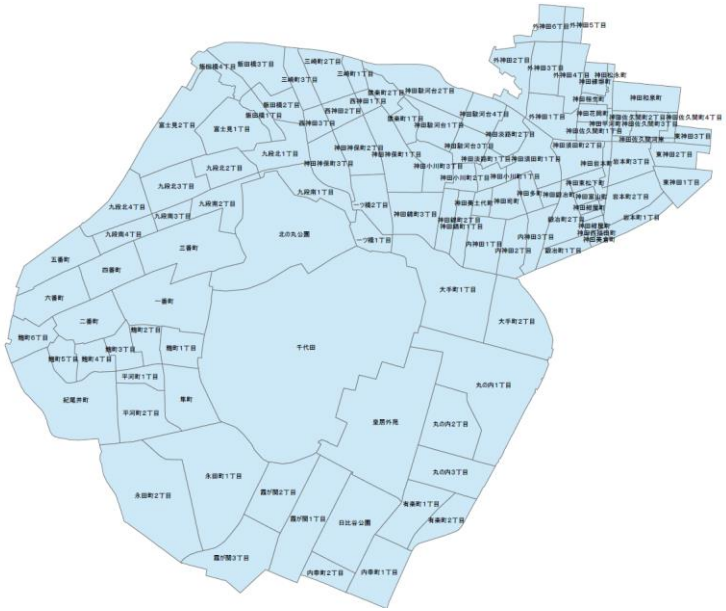
8分の1地域メッシュ（約125m）
4分の1地域メッシュを縦横2等分

データ名	出所
地理院タイル	国土地理院『電子地形図（タイル）』

町丁・字等別集計と地域メッシュ統計



212メッシュ（4分の1地域メッシュ）



116町丁・字等

（資料）「国勢調査」（総務省統計局）（地域メッシュ枠は2分の1地域メッシュ枠を加工して作成）

○東京都千代田区における4分の1地域メッシュの区画と町丁・字等別の境域を示す。
○地域メッシュの区画は、等分に区切られているため統計値の比較を行いやすいが、各地域メッシュ区画に付与された地域メッシュコードが10ケタの数値であるため（例：5339463022）、当該区画の位置や範囲は、GISを用いて確認する必要がある。
○一方の町丁・字等の境域は、大きさが地域で異なるため、統計値の比較等を行う際に配慮する必要があるが、地域名は一般的に使われる町丁字名と概ね一致しているため（例：神田神保町3丁目）、分析結果を理解しやすい。

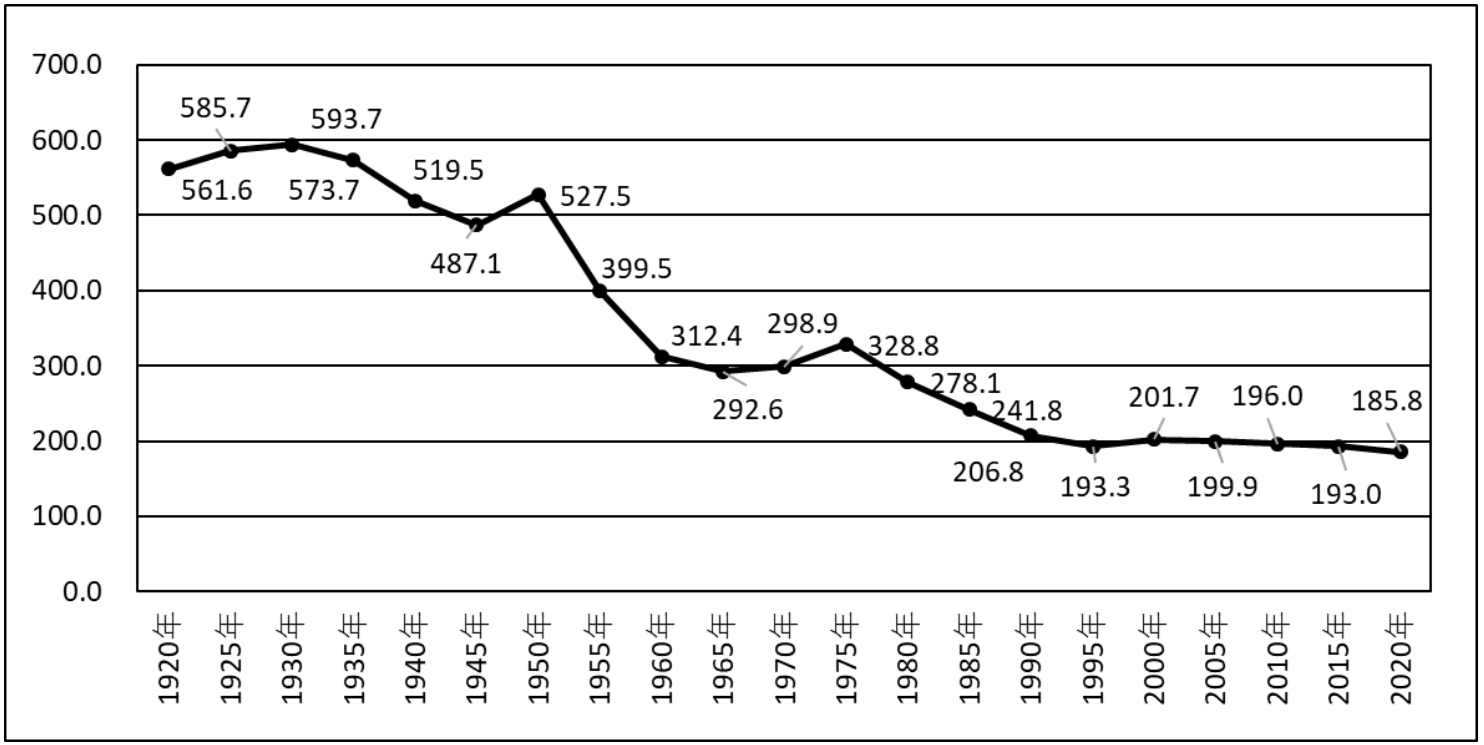
出生力の指標と標準化女性子ども比

出生力の指標

- 普通出生率
  - 人口1000人あたりの年間出生数
  - 令和2年
    - 出生児数：840,835人
    - 日本人人口：121,541,155人
    - 普通出生率＝6.92‰
- 女性子ども比（CWR：**C**hild **W**oman **R**atio）
  - 15～49歳の女性人口1000人当たりの0～4歳人口
  - 令和２年（人口総数で計算）
    - 0～４歳人口：4,516,082人
    - 15～49歳女性人口：24,299,934人
    - 女性子ども比＝185.8 ‰

（資料）人口動態調査、国勢調査

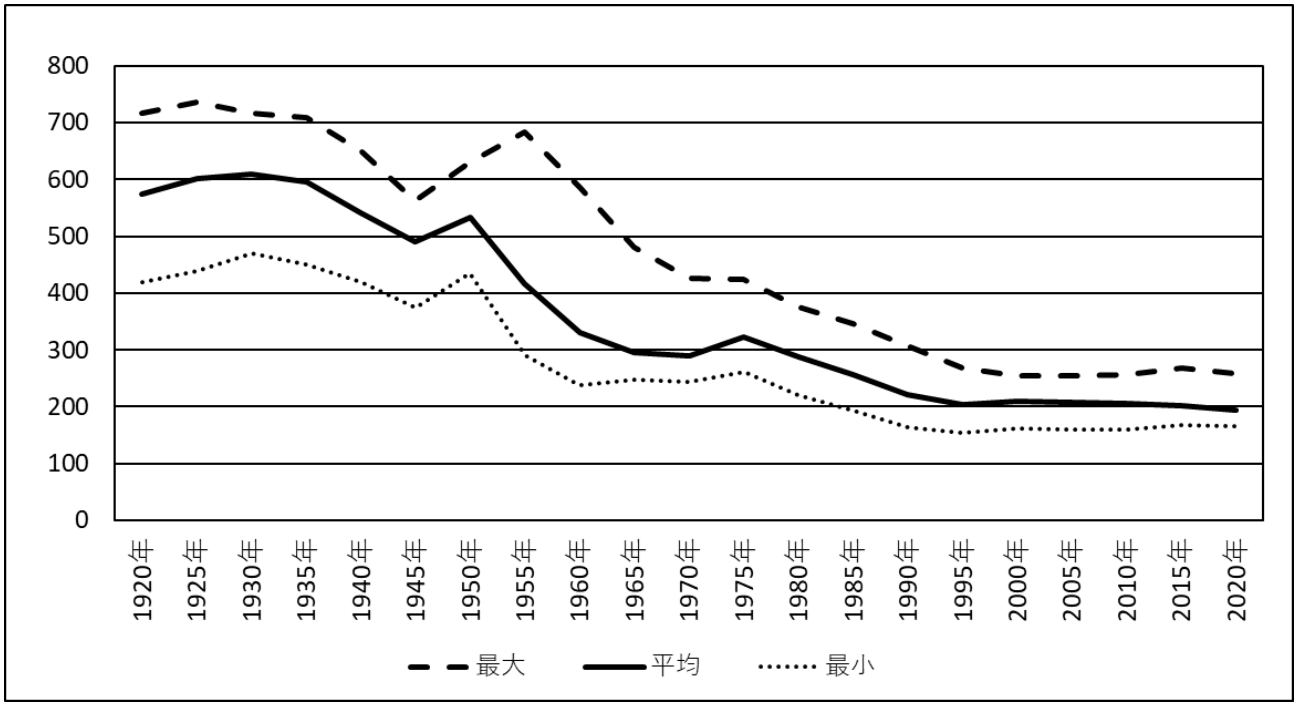
女性子ども比の推移（1920～2020年）



〇1920年～1940年の女性子ども比は、500.0以上と高い値を示す。1945年は終戦の年で、女性子ども比は487.1と500.0よりも低くなるが、5年後の1950年に527.5と上昇した。その後1965年に292.6となるまで下降傾向を示している。
〇その後1975年に328.8と再び上昇したが、その後は下降傾向となり1990年以降は200.0前後で推移している。
〇直近の2020年の女性子ども比は185.8となり、1920年からの100年間で最も低くなった。

（資料）国勢調査（総務省統計局）

都道府県別女性子ども比の推移



〇1920年～2020年の都道府県別の女性子ども比の平均、最大、最小のグラフで、平均は全国と同様の傾向を示す。
〇最大と最小の差は1975年頃まで大きく、1980年以降は小さくなっている。
〇2020年における都道府県別の女性子ども比の最大と最小の値を確認すると、最大は沖縄県の258.0、最小は東京都の165.6である。

女性子ども比の地域比較に当たっては、地域の年齢構造に留意する必要がある。

一般に、出生力の強さは年齢で異なることから、比較する地域における15～49歳女性人口が同じ場合でも、20歳代が多い地域と40歳代が多い地域では地域の出生力は異なる。女性子ども比は各地域の年齢構造の影響を受ける。

地域の年齢構造の影響を取り除いて出生力を比較するためには、**標準化**という処理が行われる。

（資料）国勢調査（総務省統計局）

標準化女性子ども比の計算

標準化の定義

- 標準化法とは
  - 人口構造の差を除いて比較するために考えられた方法である。
  - 出生率や死亡率などを年齢別に計算すると、年齢による差が大きい。このような人口1000人当たりの発生数は、**年齢構造**と年齢別の**発生率**によってその水準が決まる。
    - 例えば、都市と農村のように年齢構造が異なる人口の普通率を比較する場合に、2つの比率の差は年齢構造の違いを反映していることになる。
- 任意標準人口標準化法
  - **直接法**は、比較すべき人口の男女、年齢別出生率を標準人口の年齢別人口に適用して標準人口の出生率を計算し、出生率を比較する。
  - **間接法**は、標準人口の年齢別出生率を用いて、比較しようとする人口の年齢別人口に適用して**指標出生率**を計算し、標準人口の**標準出生率**との比（**標準化係数**）を計算して、比較すべき人口の出生率に掛ける。

標準化の定義

- 2015年の全国の人口を標準人口とする場合
  - 2015年の都道府県別の標準化女性子ども比（以下「sCWR」）を間接標準化の考え方で計算する。
- **標準CWR**
  - 標準人口によるCWR（①式）

$$\text{標準CWR} = \frac{\sum_{j=15}^{45} P_{j\sim j+4}^{f} \times \alpha_{j\sim j+4}}{\sum_{j=15}^{45} P_{j\sim j+4}^{f}} \qquad ①$$

  - 分母は15~49歳の女性人口
  - 分子の$\alpha_{j\sim j+4}$は標準人口の2011～2015年の母の年齢5歳階級別出生数を2015年の年齢5歳階級別女性人口で除した値（女性の年齢別出生率に相当）
  - $\alpha_{j\sim j+4}$を「標準出生率」とする。

標準化の定義

- **指標CWR**
  - 各地域(i)の15～49歳の年齢5歳階級別女性人口に年齢5歳階級別の標準出生率を乗じて、各地域の15～49歳の年齢5歳階級別女性人口で割った値

$$\text{指標CWR} = \frac{\sum_{j=15}^{45} P_{j\sim j+4}^{if} \times \alpha_{j\sim j+4}}{\sum_{j=15}^{45} P_{j\sim j+4}^{if}} \qquad ②$$

  - 指標CWRは標準出生率を適用した各地域の年齢構造によるCWR

標準化の定義

- **標準化係数**
  - 標準化係数（＝標準CWR／指標CWR）は、各地域の年齢構造に対する標準人口の年齢構造の比を表す

- **標準化女性子ども比**（sCWR）
  - 各地域のCWRにこの標準化係数を乗ずることによって計算できる。

$$\text{標準化CWR} = \frac{\text{標準CWR}}{\text{指標CWR}} \times CWR \qquad ③$$

標準化の定義

- **標準化女性子ども比**
  - 標準CWRをCWR^I、標準化CWRを $sCWR^I$、各地域のCWRをCWR^{i}とすると、 sCWR^{i}は以下のように④式で表せる

$$\mathrm{sCWR}^i = \frac{CWR^I}{\text{指標CWR}} \times \mathrm{CWR}^i$$

$$= CWR^I \times \frac{\sum_{j=15}^{45} P_{j\sim j+4}^{if}}{\sum_{j=15}^{45} P_{j\sim j+4}^{if} \times \alpha_{j\sim j+4}} \times \frac{P_{0-4}^{i}}{\sum_{i=15}^{45} P_{j\sim j+4}^{if}}$$

$$= CWR^I \times \frac{P_{0-4}^{i}}{\sum_{i=15}^{45} P_{j\sim j+4}^{if} \times \alpha_{j\sim j+4}} \qquad ④$$

標準化の計算手順

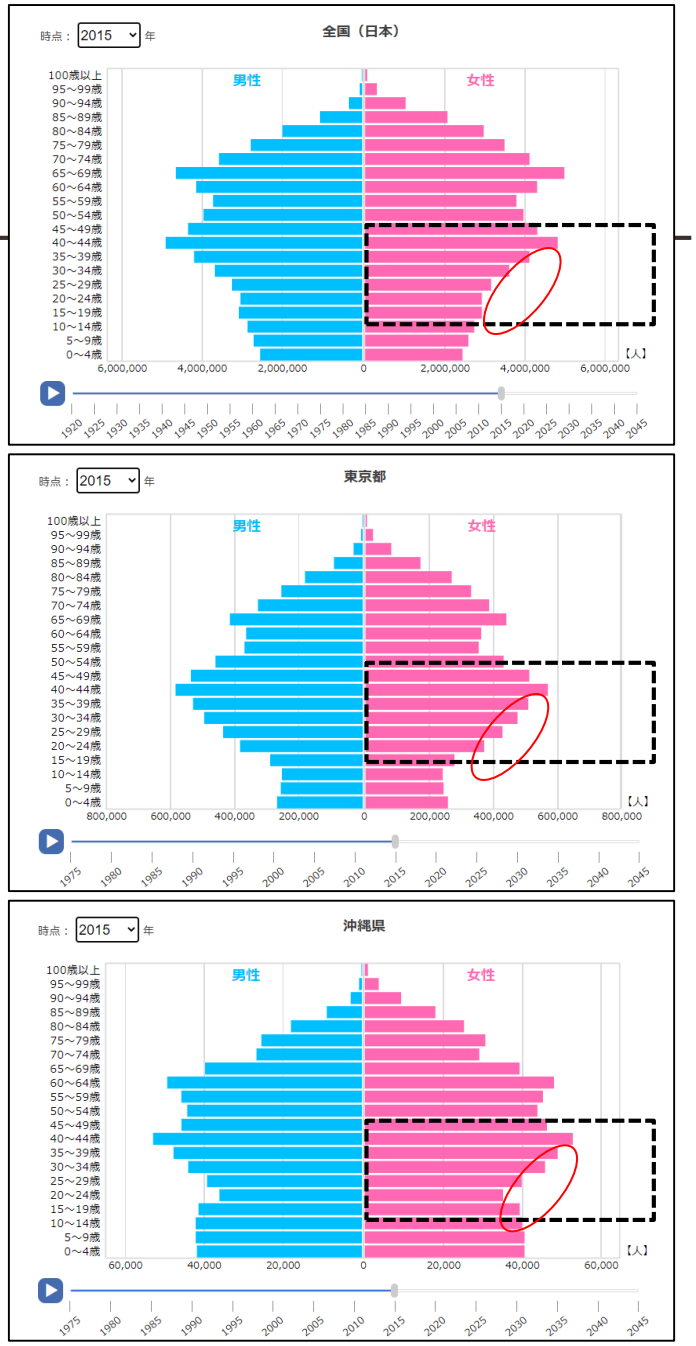
- 2015年の東京都、沖縄県の計算結果

	標準出生率（1）	東京都		沖縄県	
		女性人口（2）	（1）×（2）（3）	女性人口（2）	（1）×（2）（3）
15～19歳	21.8	277,599	6,055	39,329	858
20～24歳	158.2	370,534	58,614	35,076	5,549
25～29歳	445.7	427,543	190,567	39,813	17,746
30～34歳	507.7	474,637	240,975	45,741	23,223
35～39歳	275.0	508,347	139,784	48,942	13,458
40～44歳	47.4	569,560	26,974	52,784	2,500
45～49歳	1.2	511,490	626	46,240	57
計	198.4	3,139,710	663,594	307,925	63,389
		524,939（0～4歳人口）	0.1672（CWR）	82,414（0～4歳人口）	267.6（CWR）
	指標CWR	663,594/3,139,710	0.2114	63,389/307,925	205.9
	標準化係数	198.4/211.4	0.9386	198.4/205.9	0.9637
	sCWR	167.2×0.9386	0.1569	267.6×0.9637	257.9

（資料）人口動態調査、国勢調査

標準化CWRの結果

- 東京都と沖縄県の差
  - 東京都、沖縄県の女性子ども比は167.2、267.6、標準化女性子ども比は156.9、257.9 となっており、東京都、沖縄県ともに標準化により女性子ども比は低くなる。
  - 標準化女性子ども比は年齢構造の影響を取り除いた出生力を示し、標準化により沖縄県と東京都の出生力の差は100.4から101.0と拡大した。



標本調査の基礎と単純無作為抽出の概要

標本調査理論

- 標本調査理論は、標本調査を行うための統計理論で、大きく分けて**標本抽出**、**推定**、**誤差評価**という三つの方法論からなる。
  - 例：企業の一社当たり売上高平均を調べる場合
  - **標本抽出**
    - 対象とする企業全体（母集団）から実際に調査する企業（標本）を選びだす。
  - **推定**
    - 抽出した標本企業の情報から母集団企業における一社当たり売上高平均を推測する。推定の方法と標本抽出の方法は関連しており、推定は標本の抽出方法を考慮しながら行う。
  - **誤差評価**
    - 結果の精度あるいは誤差の大きさを見積もる。

標本調査は、誤差が生じることを容認した上で、企業全体の売上高平均を知る方法である。結果を正しく解釈するには、誤差の大きさを評価しておく必要がある。

母集団

- 母集団と要素
  - 調査対象全てからなる集団を母集団といい、Uで表す。
  - 調査対象一つ一つを要素という。
- 母集団サイズ
  - 母集団に含まれる要素の数を母集団サイズと呼び、N で表す。
- 変数
  - 変数とは、個々の要素に応じて変わる特性のことである。目的とする変数をyで表し、第i要素の変数値を y_i で表す。

母集団特性値

- 母集団特性値
  - 企業の売上高の総計や平均などの母集団の変数値$y_i, \cdots, y_N$を要約した値を母集団特性値という。
  - 母集団特性値を一般にθで表す。統計調査の目的はθを知ることである。各企業の売上高など個々のy_iを知ることが目的ではない。
  - 例えば、母集団総計を知ることが目的であれば、一般的な表記におけるθをT_yで置き換えればよい。

$$\text{母集団総計：} T_y = \sum_U y_i$$

$$\text{母集団平均：} \mu_y = \frac{1}{N} T_y$$

$$\text{母集団分散：} \sigma_y^2 = \frac{1}{N-1} \sum_U (y_i - \mu_y)^2$$

$$\text{母集団標準偏差：} \sigma_y = \sqrt{\sigma_y^2}$$

標本抽出

- 調査の目的は母集団特性値θを知ることである。
- 全数調査
  - 母集団の全ての要素の変数値を調べることで母集団特性値θを把握できる
- 標本調査
  - 母集団から抽出した標本だけを調べる調査である。
  - 抽出した標本をsで表す。標本sに含まれる要素の数を標本サイズという。標本サイズをnで表す。
- 抽出単位
  - 標本を抽出するときには、個々の要素を直に選びだすこともあれば、要素のまとまりを単位として選び出すこともある。
  - 標本を選ぶときの単位を抽出単位と呼び、要素のまとまりを集落という。

無作為抽出法

- 確率抽出法と比確率抽出法
  - 全ての可能な標本から標本を選ぶ方法には大きく分けて確率抽出法と非確率抽出法がある。
  - 確率抽出法あるいは無作為抽出法では、各標本に対して選ばれる確率$p_{(t)}$を与え、この確率に従って標本を選び出す。
  - $p_{(t)}$は非負の実数であり、その合計は1である。
  - 確率抽出法によって選ばれた標本を無作為標本という。
  - また、確率$p_{(t)}$の与え方を標本抽出デザインと呼ぶ。

統計量と推定量

- 統計量
  - 標本sに含まれる要素の変数値y_iから計算される量を統計量と呼ぶ。
  - 統計量の例

$$\text{標本総計：} \sum_s y_i$$

$$\text{標本平均：} \bar{y} = \frac{1}{n}\sum_s y_i$$

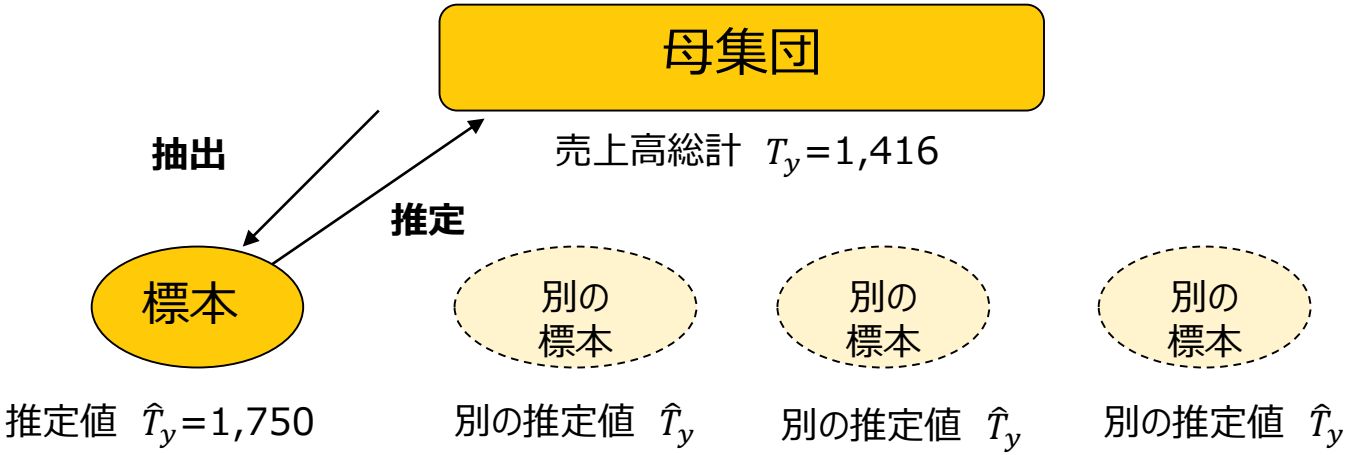
$$\text{標本分散：} S_y^2 = \frac{1}{n-1}\sum_s (y_i - \bar{y})^2$$

- 推定量
  - 調査の目的は母集団特性値θを知ることである。
  - 標本の変数値y_iから母集団特性値θを推測することを推定という。
  - またθを推定する具体的な方法・計算式をθの推定量と呼ぶ。
  - 具体的変数値を推定量に当てはめ計算した結果、得られた数値をθの推定値と呼ぶ。

誤差評価

- 誤差
  - 調査の目的は母集団特性値θを知ることであるが、普通推定値$\hat{\theta}$と真のθとは完全には一致しない。両者の差$\hat{\theta} - \theta$を誤差という。
  - 調査結果を適切に利用するためには、このような誤差の大きさを見積もっておくことが不可欠である。
    - 例えば継続調査（時系列比較）では、誤差に過ぎない推定値の変化を実質的な変化と見誤らないようにしなければならない。
- 標本誤差
  - 誤差には標本誤差と非標本誤差がある。
  - 標本誤差は、標本だけを調べることから生じる誤差である。確率標本では、標本誤差の大きさを理論的に見積もることができる
  - 非標本誤差は記入ミスなど標本誤差以外の誤差をいう。

標本誤差の評価

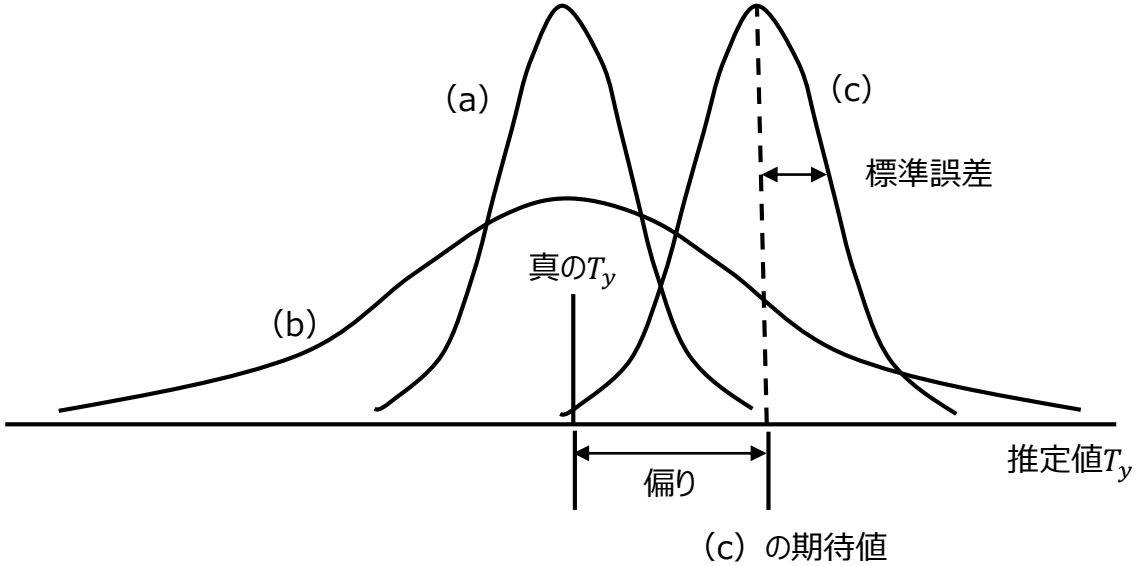


上図の場合、誤差は1750-1416＝334と計算できるが、実際には真の値T_yは分からないので、誤差の大きさを直接求めることはできない。

標本調査を何度も繰り返したとき、推定値がどの程度変動するのかを調べ、変動の大きさで誤差を評価する。

誤差を評価するには、標本調査を繰返したときに得られる推定値の分布の様子を調べる必要がある。

標本誤差の評価



推定値の分布が（a）の場合：標本調査を繰返しても推定値$\hat{T}_y$は大きく変わらない
推定値の分布が（b）の場合：推定値が真のT_yに近いことあるが大きく離れることもある
推定値の分布が（c）の場合：分布（a）と推定値の変動の大きさは同じであるが得られる推定値のほとんどは真のT_yよりも大きい

推定量の誤差の大きさは、このような推定値の分布の様子を指標化することで評価する

推定量の分散・標準誤差

- 推定量の分散
  - $V(\hat{\theta}) = \sum_s p_{(t)} \{\hat{\theta}_{(t)} - E(\hat{\theta})\}^2 = E\left[\{\hat{\theta}_{(t)} - E(\hat{\theta})\}^2\right]$
- 推定量の標準誤差
  - $SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}$

期待値$E(\hat{\theta})$：標本調査を繰返したときの推定値の“平均”
標準誤差$SE(\hat{\theta})$：標本調査を繰返したときの推定値の“標準偏差”

$V(\hat{\theta})$と$SE(\hat{\theta})$はいずれも推定値$E(\hat{\theta})$の周辺でどのくらいバラツクのかを表す指標

したがって、$V(\hat{\theta})$や$SE(\hat{\theta})$が小さい推定量は精度あるいは信頼性が高いといえ、推定量$\hat{\theta}$の誤差の大きさを評価する指標としては、推定量の分散や標準誤差を用いる。

単純無作為抽出法

- 抽出ウェイトを用いた推定量の表現
  - 抽出ウェイトとは推定用のウェイトであり、単純無作為抽出法では

$$w_i = \frac{N}{n}, (i \in s)$$

である。

非復元単純無作為抽出法、復元単純無作為抽出法の母集団総計T_yの推定量は以下のとおりになる。

非復元単純無作為抽出法	復元単純無作為抽出法
$\hat{T}_y = \sum_s \breve{y}_i$	$\hat{T}_y = \sum_s \breve{y}_i$
$\hat{V}(\hat{T}_y) = (1-f)nS_{\breve{y}}^2$	$\hat{V}(\hat{T}_y) = nS_{\breve{y}}^2$

ただし、$\breve{y}_i = w_i y_i = N\frac{y_i}{n}$　であり、$S_{\breve{y}}^2$は、$\breve{y}_i$の標本分散である。

$1-f = 1-\frac{n}{N}$は有限母集団修正項という。

標本サイズの定め方

- 目標精度
  - 標本サイズn を定めるには、まず推定値の標本誤差をどの程度に抑えたいのかを決める必要がある。これを目標精度という。
    - 例えば、母集団割合p_yを推定したいとき、推定値$\hat{p}_y$は、真の値p_yから±5ポイント程度ズレてもよいのか、あるいは±1ポイントという高い精度が必要なのか、ということである。
- 信頼水準
  - 目標精度を定めて、その範囲を設定しても全ての推定値がその範囲に収まるわけではない。
  - 適当なαを定め、調査を繰返したとき得られる推定値のうち100×（１－α）%がθ ± dの範囲に入れば良いと考える。 αとしては、0.05や0.01とすることが多い。
  - 標本サイズnが大きい推定値の分布は正規分布とみなしてよく、変数値の95%は±1.96×標準偏差の範囲に入る。

標本サイズの定め方

- 100×（１－α）%の範囲
  - $\pm z_{\alpha/2}$×標準偏差と表せる。ただし、$z_{\alpha/2}$は平均0、標準偏差1の標準正規分布の上側100 α/2%点であり、 αに応じて決まる値
  - 推定値θ̂の分布の標準偏差とは標準誤差$SE(\hat{\theta})$のことであり、目標精度の幅d は　$d=z_{\alpha/2}SE(\hat{\theta})$と表せる。
  - 一般に標準誤差$SE(\hat{\theta})$は標本サイズnの関数であり、 nが大きくなるほど小さくなる。
- 標本サイズを計算する式

$$n=\frac{z_{\alpha/2}^2 N^2 \sigma_y^2}{d^2+z_{\alpha/2}^2 N \sigma_y^2}$$

この式に値を代入し、必要なnを計算する。

(※)上側100α/2%点は$z_{1-\alpha/2}$と表すのが一般的であるが、本資料では参考資料に従って$z_{\alpha/2}$と表す

デザイン効果

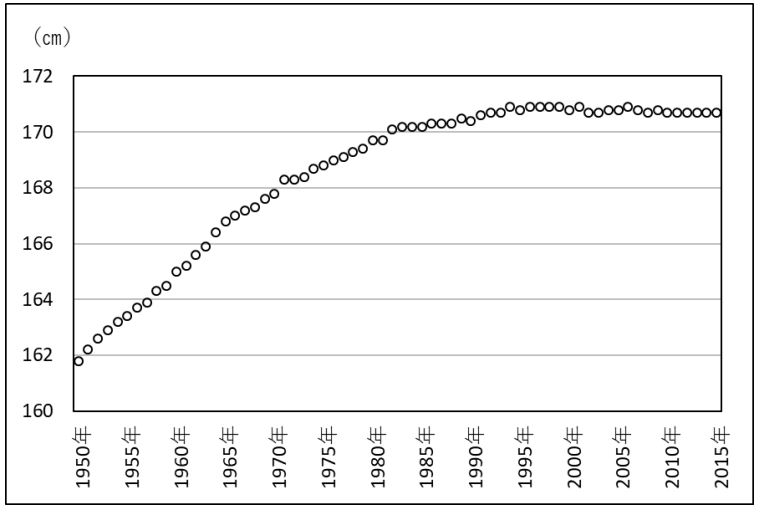
- デザイン効果とは
  - ある標本抽出デザインの“非”効率性を表す指標で、単純無作為抽出法における推定量の分散と、対象の標本抽出法の推定量の分散との比で計算される。
  - $\mathrm{Deff}=\dfrac{\text{ある標本抽出デザインにおける推定量の分散}}{\text{非復元単純無作為抽出法における推定量の分散}}$
- 複雑な標本抽出デザインの標本サイズ
  - 標本抽出デザインが複雑になれば、標本サイズを計算する式を解くのは難しくなる。
  - 標本サイズを検討する簡便な方法としてデザイン効果を利用する方法がある。デザイン効果Deffのだいたいの大きさが知られていれば、単純無作為抽出法の下で必要な標本サイズn'を求めた後に$n' = n \times$ Deffなどとすることで実際に採用する標本抽出デザインの下で必要な標本サイズnを見積もることができる。

集落抽出法の概要

集落抽出法とは

- 標本調査によって17歳男子高校生の平均身長を調べたい。
- ただし、全国の17歳男子高校生全員を掲載した名簿は存在しない。
- このため、単純無作為抽出法など、高校生個人を直接選ぶ抽出法は使えない。
- そこで、全国の高校一覧は容易に入手できるので、抽出された高校に在籍する17歳男子全員を標本とする。
- つまり、抽出単位を高校生個人とするのではなく、高校生の集団である高校とすればよい。

17歳男子高校生の平均身長の推移

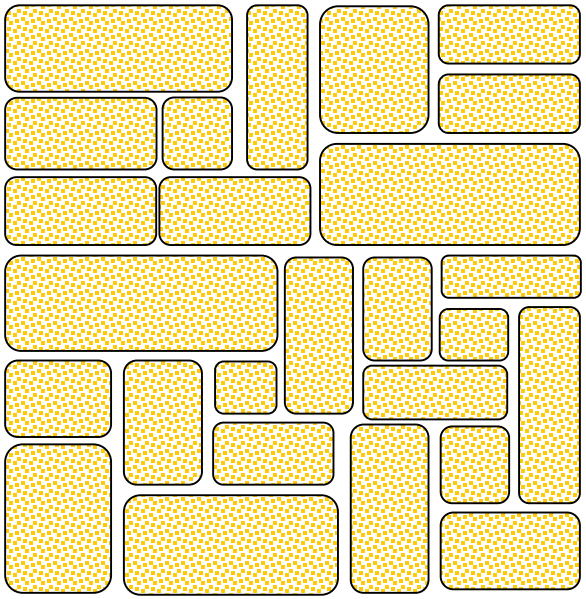


（出所）学校保健統計調査

集落抽出法とは

- 母集団をM個のグループに分割する。このグループを集落と呼ぶ
- M個の集落からm個の集落を抽出する
- 標本は選ばれたm個の集落に含まれる全ての要素である。

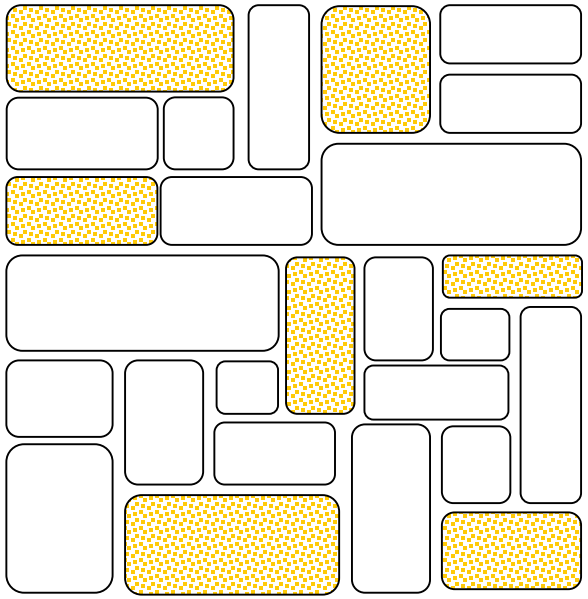
母集団U　（ N ）



M個の各集落U_a　（N_a）

抽出

標本$s = \cup_{a \in s_I} U_a$　（ $n = \sum_{a \in s_I} N_a$ ）



m個の各集落U_a　（N_a）

集落抽出の利点

1. 母集団の全要素をリストアップしなくてもよく、集落のリストさえあればよい。
   - 例：母集団を全国の全ての入院患者とするとき、全患者リストの作成は困難だが、病院のリストは容易に入手可能
2. 母集団サイズNが未知でもよい。母集団における集落の数Mさえ分かれば、標本抽出も推定も可能
3. 要素を抽出単位とするよりも低コストで標本抽出や実査を行える。
   - 地理的なまとまりを集落とすれば、調査対象が点在する場合に比べ、調査員の移動にかかるコストを抑えられる。
   - 子どもを対象に教室で一斉に調査をすれば、標本サイズを用意に大きくできる。

集落抽出の難点

1. デザイン効果が１を超えやすい
   - 同じ標本サイズの単純無作為抽出法よりも、推定量の精度は低いことが多い。その理由は同一集落内の要素は似たような性質を持ちやすく、標本に含まれる情報が重複し、ムダが増える。
   - 例えば、同じ学校の高校生どうしは、他校の高校生と比べて学力は似通っている。
   - 一校で多数の高校生を調べても学力に関する推定値は、選ばれる学校に応じて大きく変動してしまうので、推定量の精度低下を防ぐ工夫が必要。
2. 集落のサイズが不揃いのとき、標本サイズnを固定できない。
   - nは選ばれた集落によって変わる。ただし、集落サイズが全て等しいときには、抽出する集落の数を固定すれば標本サイズも固定される。

集落抽出の方法

- 母集団UをM個の集落$U_1, \ldots, U_M$に分割する

$$U = U_1 \cup U_2 \cup \cdots \cup U_M = \bigcup_{a \in U_I} U_a$$

  - $U_I = \{1,2, \ldots, M\}$は集落の集合である。添え字のIはローマ数字である。第a集落に含まれる要素の数、つまり集落サイズをN_aとする。必ずしもN_aが等しい必要はない。母集団サイズNは集落サイズ$N_1, \ldots, N_M$の合計$N = \sum_{a \in U_I} N_a$ である。
  - 集落総計$T_{y,a}$や集落平均$\mu_{y,a}$、集落分散$\sigma^2_{y,a}$はそれぞれ第a集落含まれる全要素から求めた総計や平均、分散である。

$$T_{y,a} = \sum_{U_a} y_i \ , \ \ \mu_{y,a} = \frac{1}{N_a} T_{y,a} \ , \ \sigma^2_{y,a} = \frac{1}{N_a - 1} \sum_{U_a} \left(y_i - \mu_{y,a}\right)^2 \ , (a \in U_I)$$

  - 母集団総計T_yは集落総計$T_{y,1}, \ldots, T_{y,M}$の合計である。

$$T_y = T_{y,1} + \cdots + T_{y,M} = \sum_{a \in U_I} T_{y,a}$$

集落抽出の方法

- 集落を単位としてm個の集落を単純無作為抽出する。
  - 集落の抽出率を$f_I = {}^{m}/_{M}$とする。
  - 抽出された集落の集合をs_Iとする。例えば集落1と4が抽出されれば$s_I = \{1,4\}$ である。
  - 抽出された集落に属する全ての要素を標本sとする。

$$s = \bigcup_{a \in s_I} U_a$$

標本サイズnは抽出された集落サイズの合計である。

$$n = \sum_{a \in s_I} N_a$$

母集団総計T_yの推定

- 母集団総計T_yの推定は、単純無作為抽出の“要素”を“集落”と読み替えて考えていく。つまり、第i要素の変数値y_iの代わりに第a集落の集落総計$T_{y,a}$を用いればよい。
- したがって母集団総計T_yの線形推定量は、要素を抽出単位としたときの線形推定量$\hat{T}_y = \sum_s w_i y_i$ のy_iを$T_{y,a}$で置き換えればよい。抽出ウェイトもw_aも集落単位として考える。例えば集落を単純無作為抽出したのであれば、$w_a = {}^{M}/_{m}$ である。

$$\hat{T}_y = \sum_{a \in s_I} w_a T_{y,a} = \sum_{a \in s_I} w_a \sum_{i \in U_a} y_i = \sum_s w_a y_i$$

- 集落総計$T_{y,a}$さえ分かれば、集落内の個々の要素の変数値y_iは不要である。最終的に$\hat{T}_y$は,抽出ウェイトw_aによる各要素の加重変数値の標本総計となる。

$\widehat{T_y}$の分散やその推定量の推定

- T_yの分散やその推定量を求めるときにも、これまでのy_iを$T_{y,a}$で読み替える。例えば変数y_iの母集団分散 σ_y^2や標本分散 S_y^2は、集落総計$T_{y,a}$の母集団における分散 $\sigma_{T_y}^2$ や標本における分散 $S_{T_y}^2$ で置き換える。

$$\sigma_{T_y}^2 = \frac{1}{M-1}\sum_{a\in U_I}\left(T_{y,a} - \frac{1}{M}\sum_{a\in U_I} T_{y,a}\right)^2$$

$$S_{T_y}^2 = \frac{1}{m-1}\sum_{a\in s_I}\left(T_{y,a} - \frac{1}{m}\sum_{a\in s_I} T_{y,a}\right)^2$$

- 母集団平均や母集団割合など母集団総計T_yを利用する特性値を推定するときには、集落総計$T_{y,a}$さえわかればよい。

単純無作為集落抽出法

- 集落を非復元単純無作為抽出したときの推定量
- 第a集落の抽出ウェイトw_aは、M個の集落からm個の集落を選ぶので、$w_a = {}^{M}/_{m}$である。
- 母集団総計の線形推定量とその分散とその推定量

$$\hat{T}_y = \sum_{a \in s_I} w_a T_{y,a} = {}^{M}/_{m} \sum_{a \in s_I} T_{y,a} = {}^{M}/_{m} \sum_{s} y_i$$

$$V(\hat{T}_y) = M^2(1 - f_I)\frac{1}{m}\sigma_{T_y}^2 = \frac{M(M-m)}{m(M-1)} \sum_{a \in U_I} \left(T_{y,a} - \frac{1}{\mathrm{M}} T_y \right)^2$$

$$\hat{V}(\hat{T}_y) = M^2(1 - f_I)\frac{1}{m} S_{T_y}^2$$

推定量の分散$V(\hat{T}_y)$は抽出する集落の数mにほぼ反比例する。

集落抽出法で推定量の分散を小さくするには、標本サイズnというよりは、抽出する集落の数mを大きくする必要がある。

単純無作為集落抽出法

- 母集団平均の線形推定量

$$\hat{\mu}_y = \frac{1}{N}\hat{T}_y = \frac{1}{N}\frac{M}{m}\sum_{a\in s_I} T_{y,a} = \frac{1}{N}\frac{M}{m}\sum_s y_i$$

  - 要素を単純無作為抽出すると、線形推定量$\hat{\mu}_y = \frac{1}{N}\hat{T}_y$は標本平均$\bar{y}$に一致するが、集落を単純無作為抽出したときは一致しない。

- 母集団総計の比推定量
  - 一般に集落総計$T_{y,a} = \sum_{U_a} y_i$は集落サイズN_aと相関が高いことが多い。
    - 学校の在籍数N_aが多いほど、生徒の身長総計も大きくなる
  - サイズを用いた比推定量$\hat{T}_{y,N}$は線形推定量$\hat{T}_y$よりも精度が高いと期待できる。

$$\hat{T}_{y,N} = N\frac{\hat{T}_y}{\hat{N}} = N\frac{M}{m}\sum_{a\in s_I} T_{y,a} \Big/ \frac{M}{m}\sum_{a\in s_I} N_a = \frac{N}{n}\sum_s y_i = N\bar{y}$$

  - 母集団総計の比推定量$\hat{T}_{y,N}$は、標本平均$\bar{y}$を母集団サイズNで拡大したものとなる。
  - そのため、母集団平均$\hat{\mu}_y$の推定量$\hat{\mu}_{y,N} = \frac{\hat{T}_{y,N}}{N}$は標本平均$\bar{y}$に一致する。

比推定量

- 比推定量とは
  - 表の例で考える。調査の目的は企業の売上高yの母集団総計T_yを知ることである。
  - 非復元単純無作為抽出標本に基づく推定値として$\hat{T}_y$ =8,187が得られたとする。ここで、補助変数である資本金xがあらかじめわかっているものとする。標本からあえて推定したところ、$\hat{T}_x$ =793が得られたとする。真の資本金総計663は、この推定値の84%に当たる。したがって売上高yについても母集団総計T_yの推定値を$\hat{T}_y$ =8,187の84%としてはどうか。
  - 一般に、二つの母集団総計の線形推定量$\hat{T}_y$と$\hat{T}_x$の比を利用した推定量を比推定量という。

	母集団総計	線形推定値
売上高y	T_y =???	$\hat{T}_y$ =8,187
資本金x	T_x =663	$\hat{T}_x$ =793

比推定量の性質とリサンプリング

- 比推定量の性質
  - 比推定量の分散$V\left(\hat{T}_{y,R}\right)$およびその推定量$\hat{V}\left(\hat{T}_{y,R}\right)$は、一般に理論式によって正確に表現することはできない。
  - また一般に、推定量$\hat{\theta}$がより複雑になると、単純無作為抽出法であっても$V\left(\hat{\theta}\right)$や$\hat{V}\left(\hat{\theta}\right)$の正確な理論式は導出が困難となる。
  - そういった複雑な$\hat{\theta}$を推定するには、大きく二つの方法がある。
    1. 推定量$\hat{\theta}$を線形近似することで$\hat{V}\left(\hat{\theta}\right)$を近似する方法
    2. 標本の分割、あるいは標本からの再抽出（リサンプリング）を行い、いわばシミュレーションで求める方法である。

ブートストラップ法

- ブートストラップ法では、無作為な副標本を標本sからB個独立に抽出する。
- 推定量$\hat{\theta}$の分散$V(\hat{\theta})$に対するブートストラップ推定量は、

$$\hat{V}_{\mathrm{b}}(\hat{\theta}) = E_{\hat{F}}\{\hat{\theta}^* - E_{\hat{F}}(\hat{\theta}^*)\}^2$$
$$= E_{\hat{F}}[\hat{\theta}(Y_1^*, \ldots, Y_n^*) - E_{\hat{F}}\{\hat{\theta}(Y_1^*, \ldots, Y_n^*)\}]^2$$

によって定義される。この式の計算は一般に次のモンテカルロ近似に頼らなければならない場合が多い。

$$\hat{V}_{\mathrm{b}}(\hat{\theta}) \approx \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\theta}^*)^2 = \frac{1}{B-1}\sum_{b=1}^{B}[\hat{\theta}(Y_1^{*b}, \ldots, Y_n^{*b}) - \bar{\theta}^*]^2$$

ただし、$Y_1^{*b}, \ldots, Y_n^{*b}$は、$b$回目のリサンプリングで得られたブートストラップ標本であり、$\hat{\theta}^{*b}$はこれに基づくブートストラップ統計量である。また、$\bar{\theta}^*$はB個のブートストラップ統計量の平均を表している。

地域メッシュ統計を利用した集落抽出における平均の推定とその分散の漸近性

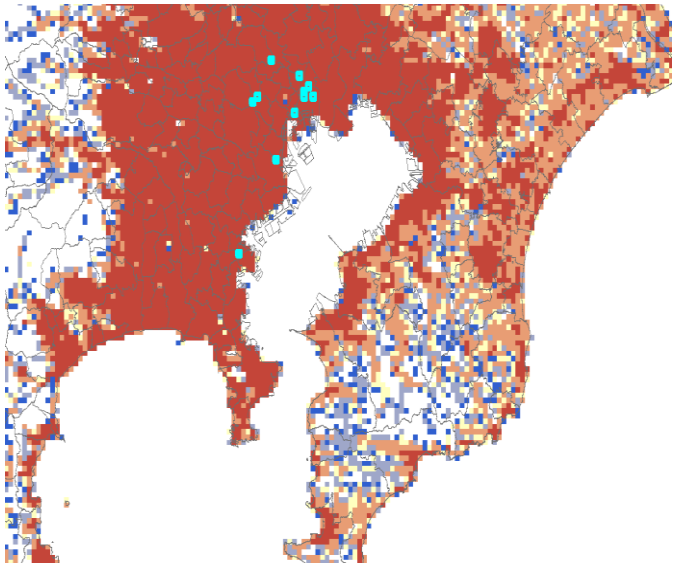
分析方法

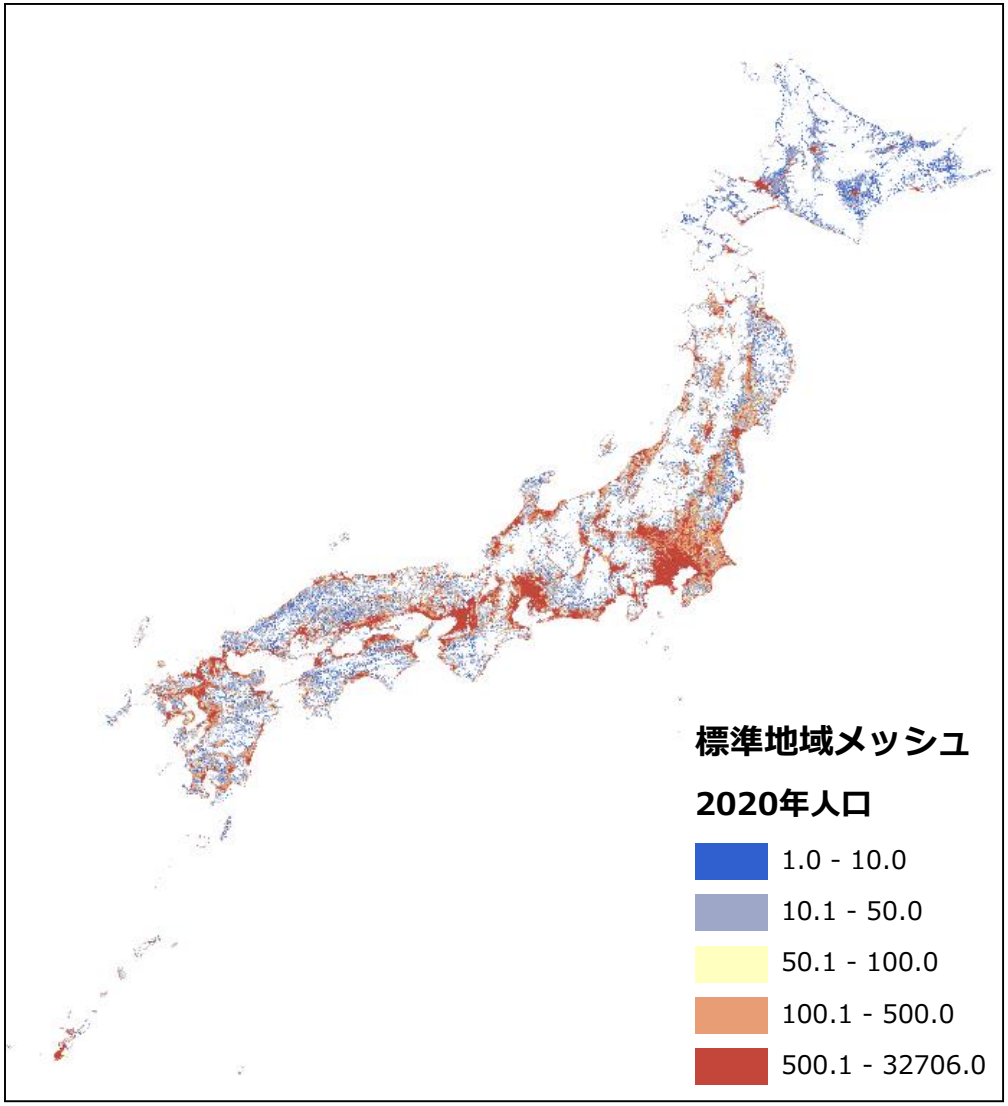
- 利用データ
  - 令和2年国勢調査地域メッシュ統計　標準地域メッシュ別データ
- 分析手順
  - 全国の標準地域メッシュ別（以下「メッシュ別」という）の0~4歳人口と15~49歳女性人口から標準化女性子ども比を算出する。
  - 標準化女性子ども比の基本統計量、ヒストグラムを確認する。
  - 女性子ども比の単純無作為抽出による標本から平均を推定し、推定結果の収束状況をグラフ化する。
  - 地域メッシュ枠の重心点データを作成し、分析対象メッシュのみ保存する。
  - 単純無作為抽出の目標精度、信頼水準を定め、標本サイズを決める。
  - 集落抽出における集落サイズと要素サイズの関係を把握する。
  - 集落抽出の標本サイズを検討し、平均の線形推定、比推定、ブートストラップ推定を行う。
  - デザイン効果を計算し、推定精度、誤差について確認する。
  - ブートストラップ分散の推定の反復回数を増やし、収束傾向を可視化する。

2020年標準地域メッシュ別人口

基本統計量

メッシュ数	176,962
平均	712.8
標準偏差	2,102.4
最小値	1
最大値	32,706



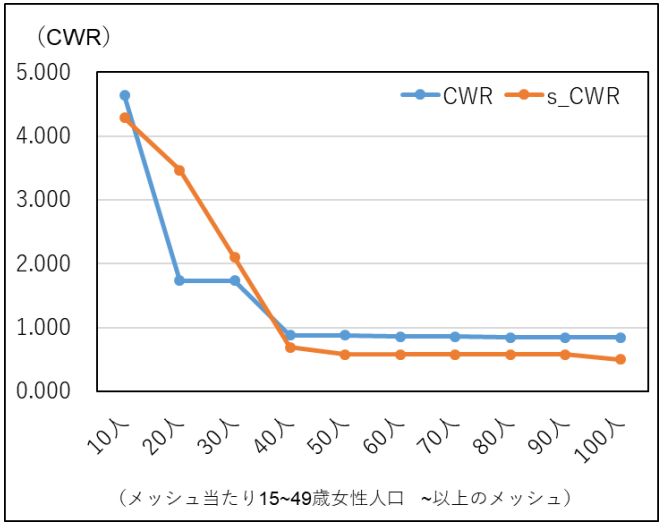


女性子ども比の基本統計量

15~49歳女性人口	変数	N	平均	標準偏差	最小値	最大値
10人以上	Woman	86948	277.01	623.61	10.00	9680.00
	CWR		0.18	0.10	0.00	4.64
	s_CWR		0.19	0.10	0.00	4.29
20人以上	Woman	67490	352.88	689.41	20.00	9680.00
	CWR		0.18	0.08	0.00	1.73
	s_CWR		0.19	0.08	0.00	3.46
30人以上	Woman	56860	414.34	734.95	30.00	9680.00
	CWR		0.19	0.08	0.00	1.73
	s_CWR		0.19	0.07	0.00	2.10
40人以上	Woman	49942	467.00	769.53	40.00	9680.00
	CWR		0.19	0.07	0.00	0.88
	s_CWR		0.19	0.06	0.00	0.69
50人以上	Woman	45125	512.12	796.42	50.00	9680.00
	CWR		0.19	0.07	0.00	0.88
	s_CWR		0.19	0.06	0.00	0.58
60人以上	Woman	41467	552.51	818.61	60.00	9680.00
	CWR		0.19	0.07	0.00	0.86
	s_CWR		0.19	0.06	0.00	0.58
70人以上	Woman	38537	589.62	837.60	70.00	9680.00
	CWR		0.19	0.07	0.00	0.84
	s_CWR		0.20	0.05	0.00	0.58
80人以上	Woman	36131	623.94	854.07	80.00	9680.00
	CWR		0.19	0.06	0.00	0.84
	s_CWR		0.20	0.05	0.00	0.58
90人以上	Woman	34122	655.71	868.46	90.00	9680.00
	CWR		0.19	0.06	0.00	0.84
	s_CWR		0.20	0.05	0.00	0.58
100人以上	Woman	32474	684.20	880.73	100.00	9680.00
	CWR		0.19	0.06	0.00	0.84
	s_CWR		0.20	0.05	0.00	0.51

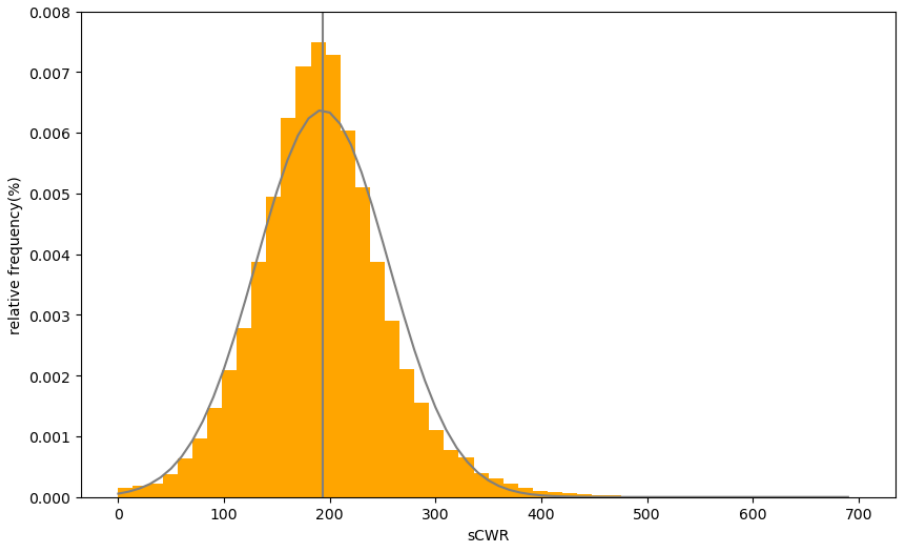
〇女性子ども比について基本統計量を算出し、分母となる15～49歳女性人口が少ないメッシュを除外して基本統計量を算出。
〇女性子ども比、標準化女性子ども比について最大値を計算すると、15～49歳女性人口が少ないメッシュでは、最大値が極端に大きな値となり安定しない。
〇15～49歳女性人口が40人以上のメッシュを分析対象とする。

15~49歳以上女性人口が~人以上のメッシュのCWR、s_CWRの最大値



女性子ども比のヒストグラム

- 女性人口が40人以上のメッシュ

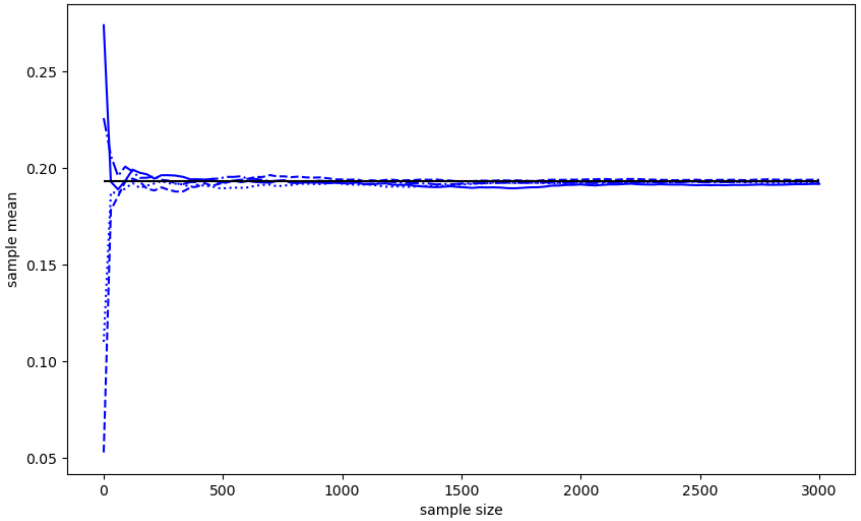


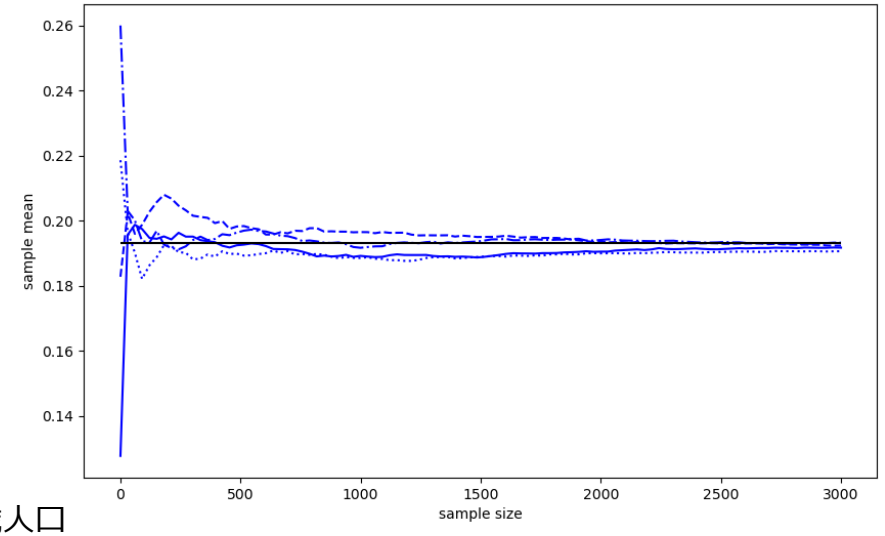
〇15～49歳女性人口が40人以上のメッシュについてヒストグラムを作成した。平均193.0を中心とした分布である。
〇平均を中心として概ね左右対称の形状であるが、右側に裾が長く、顕著に高い標準化女性子ども比となるメッシュも存在する。
〇平均193.0、標準偏差62.6の正規分布を重ねると、平均付近のメッシュの度数が多い。

変数	*N*	平均	標準偏差	最小値	最大値
Y20_POPA	49942	2364.2	3442.2	23	32706
woman	49942	467.0	769.5	40	9680
child	49942	87.1	137.7	0	1690
CWR	49942	187.1	71.3	0	882.4
s_CWR	49942	193.0	62.6	0	689.3

単純無作為抽出による平均の推定

- 単純無作為抽出
  - 女性人口が40人以上のメッシュを母集団とし、標本サイズ3000の単純無作為抽出を行った。
  - 100、200、…、3000と100区切りごとに平均を計算し、グラフ化した。
  - 標本サイズが1000程度で母平均に収束する場合もあれば3000程度で収束する場合もあることがわかる。



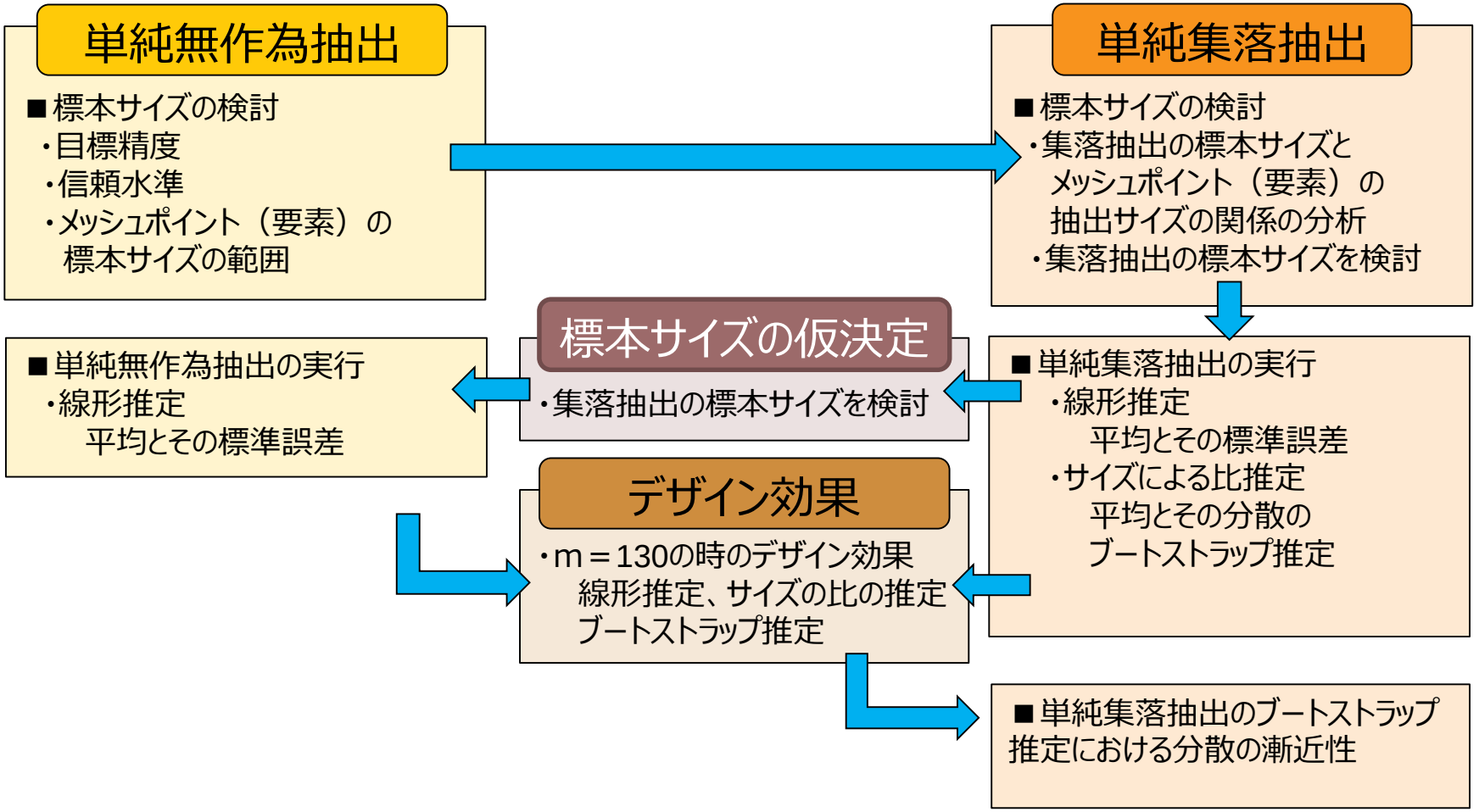


（注）グラフの縦軸の単位は15~49歳女性人口当たりの0~4歳人口

分析対象データの下処理

- 抽出を行うための下準備
  - 市区町村別の境界を「集落」、基準地域メッシュを「要素」として集落抽出を行う。
  - 市区町村境界データ
    - 令和2年国勢調査町丁・字等別境界データを市区町村コードを利用してディゾルブし、全国の市区町村境界データを作成する。
  - 地域メッシュ統計データ
    - 基準地域メッシュの重心点のポイントデータを作成し、標準化女性子ども比のデータを格納した。
    - 市区町村境界データと重心点ポイントデータを空間結合により紐づける。重心点が市区町村境界内にない場合は分析対象から除外する。
    - 標準化女性子ども比が格納されたポイントデータが市区町村内に5ポイント以上存在する市区町村を分析対象とする。
    - 対象市区町村：1507、対象メッシュ：48080

分析の枠組み


単純無作為抽出
■標本サイズの検討
・目標精度
・信頼水準
・メッシュポイント（要素）の
標本サイズの範囲
単純集落抽出
■標本サイズの検討
・集落抽出の標本サイズと
メッシュポイント（要素）の
抽出サイズの関係の分析
・集落抽出の標本サイズを検討
標本サイズの仮決定
・集落抽出の標本サイズを検討
■単純無作為抽出の実行
・線形推定
平均とその標準誤差
■単純集落抽出の実行
・線形推定
平均とその標準誤差
・サイズによる比推定
平均とその分散の
ブートストラップ推定
デザイン効果
・m＝130の時のデザイン効果
線形推定、サイズの比の推定
ブートストラップ推定
■単純集落抽出のブートストラップ
推定における分散の漸近性

単純無作為抽出

- 標本サイズの検討
  - 目標精度
    - 2010年と2015年の標準地域メッシュ別標準化女性子ども比と母平均の差は12.0となっている。
    - 母平均の目標精度dを±1～ ±3として標本サイズnを検討する。

2010年~2015年の標準化女性子ども比の平均とその差

	メッシュ数	平均	標準偏差
2015年	34131	205.7	52.6
2010年	36122	193.7	48.6
差	-1991	12.0	4.0

  - 信頼水準
    - αを0.05とすると推定値の95%が目標精度内に入り、$\lambda = 1.96$となる。
    - αを0.01とすると推定値の99%が目標精度内に入り、$\lambda = 2.56$となる。

単純無作為抽出

- 標本サイズnの計算
  - 一般にn/N が無視できるほど小さいときには、以下の式で計算する。

$$n = \lambda^2 \hat{\sigma}^2 / d^2$$

- $\hat{\sigma}^2$は母分散の推定量であるが、本分析では母分散3823.8を使用して計算した。

目標精度別の標本サイズ

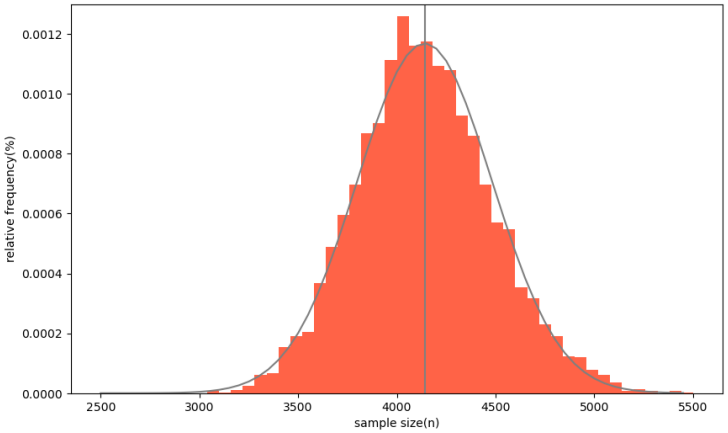
	95%水準		99%水準	
目標精度d	標本サイズn	抽出率n/N	標本サイズn	抽出率n/N
±1	14689.6	0.306	25059.8	0.521
±2	3672.4	0.076	6264.9	0.130
±3	1632.2	0.034	2784.4	0.058

- 目標精度d=±1と±2について標本サイズを比較すると、±1が±2の標本サイズの4倍となっており、抽出率が高くなる。
- 目標水準±2、95%水準の標本サイズn＝3672を参考に要素数3500~4000となる単純集落抽出の標本サイズを検討する。

単純集落抽出

- 標本サイズの検討
  - 集落標本サイズが100、110、120、130、140、150の場合について単純集落抽出をそれぞれ5000回実行し、抽出される要素（地域メッシュ）数について平均、標準偏差、最小、最大を計算した結果を表に示す。

集落サイズ	平均	標準偏差	最小	最大
100	3198.6	306.2	2262	4396
110	3514.5	311.1	2500	4734
120	3821.1	330.2	2796	5360
130	4147.9	344.2	2853	5300
140	4463.7	348.6	3241	5812
150	4776.0	360.4	3569	6123



要素（メッシュ）数のヒストグラム（m＝130）

  - 上右グラフは、集落標本サイズ130における要素数のヒストグラムである。正規分布のような形状を示している。
  - この集落標本サイズでは、要素数3500を超える標本が97%を占めるため、この標本サイズで単純集落抽出を行う。

単純集落抽出

- m=130：3パターンの集落抽出
  - 線形推定、サイズによる比推定、ブートストラップ推定を行い、誤差の評価を行った。
  - 市区町村の規模のばらつきが大きいこともあり、サイズの大きさを考慮に入れない線形推定では、誤差が大きく目標精度に収まらない。
  - ブートストラップ推定では標準誤差の推定値が小さいためデザイン効果も小さくなる。
  - 市区町村の標本サイズ130に対して、標本サイズ500の復元無作為抽出を行った。

	推定方法	平均の推定値	標準誤差の推定値	誤差（※）	目標精度	デザイン効果
① m=130 n=3741	線形推定	173.439	13.247	-19.29		185.4
	サイズによる比推定	192.288	2.350	-0.44	○	5.8
	ブートストラップ推定	192.306	1.226	-0.43	○	1.6
② m=130 n=4197	線形推定	194.060	12.956	1.33	△	189.2
	サイズによる比推定	191.775	2.102	-0.96	○	5.0
	ブートストラップ推定	191.758	1.080	-0.97	○	1.3
③ m=130 n=4846	線形推定	225.162	19.654	32.43		538.4
	サイズによる比推定	192.711	2.177	-0.02	○	6.6
	ブートストラップ推定	192.668	1.192	-0.06	○	2.0

(※)誤差＝平均の推定値　－　母平均（192.732）

単純無作為抽出

- （非復元）単純無作為抽出の実行
  - 単純集落抽出の要素数を標本サイズとして単純無作為抽出を行った。
  - 誤差は±1未満となっており、目標精度を達成している。

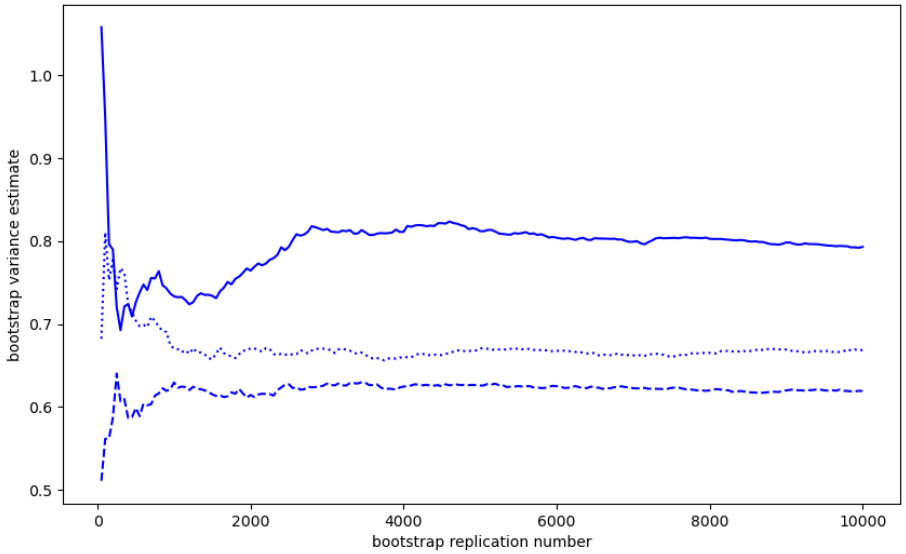
	推定方法	平均の推定値	標準誤差の推定値	誤差	目標精度
n=3741	単純無作為抽出	192.725	0.973	-0.01	○
n=4197	単純無作為抽出	192.132	0.942	-0.60	○
n=4846	単純無作為抽出	192.018	0.847	-0.71	○

単純集落抽出

- ブートストラップ推定における分散の漸近性
  - グラフは集落抽出におけるブートストラップ分散推定の漸近性について、可視化したものである。

ブートストラップ分散推定のアルゴリズム

〇3つの集落抽出（標本サイズm＝130）についてそれぞれリサンプリングを行い、平均とその分散を推定する。
〇３つの集落抽出の結果の市区町村コードのリストから無作為復元抽出を1000回行ったブートストラップ標本を構成し、その標本平均を算出する。
〇ここでは、ブートストラップ反復回数を10000として、50回ごとにブートストラップ分散推定値を計算し、グラフ化した。
〇3000~4000回ほど反復を行うと分散が収束する傾向が見られる。



まとめ

- 本分析では、抽出単位（要素）を標準地域メッシュ別データ、集落を市区町村境界とし、単純無作為抽出における推定との比較を行った。
- 集落抽出のサイズは、単純無作為抽出の目標精度から必要な要素数について検討し、その要素数以上となるような集落抽出サイズを定めた。
- 線形推定、サイズの比推定、ブートストラップ推定について誤差評価を行うと、サイズの比推定、ブートストラップ推定は目標精度内の推定が行える。
- ブートストラップ推定はサイズの比推定よりもデザイン効果が高く、この推定方法を併用することで、理論値よりも小さい集落抽出のサイズについて考察することができる。
- ブートラップ推定における分散の漸近性について可視化したところ、女性子ども比の平均の分散の推定値は収束する傾向がある。
- 集落抽出における集落のサイズ、標本サイズと精度の関係については、今後より詳しく分析する。

参考文献

- 山口喜一編著、伊藤達也・金子武治・清水浩昭（1989）『人口分析入門』古今書院
- 土屋隆裕（2009）『概説標本調査法』朝倉書店
- 金明哲編、汪金芳・桜井裕仁著（2011）『Rで学ぶデータサイエンス　ブートストラップ入門』共立出版

修正履歴

2024年4月4日：

スライド5　（国勢調査の集計地域区分→センサスの集計地域区分）

スライド7　（4分の1地域メッシュの編成範囲：政令指定都市→全国、8分の１地域メッシュの行を追加）

スライド33　（推定値T_y → 推定値$\hat{T}_y$）

スライド34　（新の値$\hat{T}_y$ → 新の値T_y ）

スライド38　（標本サイズの式の分子の$N \rightarrow N^2$）

スライド50　（母集団平均の線形推定量とその分散とその推定量→母集団平均の線形推定量）

スライド52　（「また一般的に、」を追加）