データサイエンス研究
創刊号，2022 年 3 月
63 頁～77 頁

公的統計匿名データを利用したデータサイエンス講義のための取組み

白川清美[1,2]，武藤杏里[1]，田中雅行[2]，合田智一[2]，千葉亮太[2]

Efforts for data science lectures using the official statistics anonymized microdata

Kiyomi Shirakawa[1,2], Anri Muto[1], Masayuki Tanaka[2], Tomokazu Goda[2], Ryota Chiba[2]

昨今，ビッグデータを利用したデータサイエンスが話題となっており，それに関連して公的統計の調査票情報（ミクロデータ），匿名データ及びオーダーメード集計の二次的利用による実証分析への関心が高まっている．特に，統計の作成等の範囲の拡大や手数料額の引き下げなど利用者の利便性向上に関する改正が行われた統計法の施行（2019 年 5 月）以降，情報量が多い調査票情報の利用件数が増加している．一方，調査票情報から直接識別子や間接識別子などの情報の削除や変更により作成された匿名データの利用件数が，調査票情報利用増加による影響か，増加していないようである．また，新たな需要が見込まれた「高等教育目的」の利用が，極端に少ないことの影響もあるようである．

本稿では，これまでの著者の「学術研究目的」による匿名データの利用の経験を活かした「高等教育目的」への利用拡大を図るため，その取組みの準備と講義方法を提案する．この取組みの成果として，履修者の履修目標の達成に加え，これまで匿名データの利用を控えていた教育者や研究者に対して，匿名データの利用申請や講義のための準備作業の負担を軽減することやこの講義でのデータ分析結果を分析事例として公表することによる，適用範囲の拡大と利用促進を可能とする．さらに，匿名データの利用の拡大が，新たなサービス提供へと発展し，公的統計の二次的利用における利用への利便性向上を期待する．

キーワード：PBL（Project Based Learning），ビッグデータ，実証分析，データ分析前処理，公的統計二次利用，高等教育目的

1. はじめに

昨今，ビッグデータを利用したデータサイエンスが話題となっており，それに関連して公的統計の調査票情報（ミクロデータ），匿名データ及びオーダーメード集計の二次的利用による

[1] 立正大学，[2] 一橋大学経済研究所（[1]Rissho University, [2]Institute of Economic Research, Hitotsubashi University）

実証分析への関心が高まっている．

これらの調査票情報，匿名データ及びオーダーメード集計による公的統計の二次的利用は，2009 年 4 月の統計法（2007 年法律第 53 号）の施行以降，エビデンスに基づいた実証分析が可能であるため，多くの有用な研究が発表されている．また，AI（Artificial Intelligence）の普及により，調査票情報及び匿名データを利用した機械学習など，分析に用いる数理モデルが多様化している．特に，統計の作成等の範囲の拡大や手数料額の引き下げなど利用者の利便性向上に関する改正が行われた統計法の施行（2019 年 5 月）以降，科学研究費などの競争的資金を獲得していない研究者も調査票情報が利用可能になったため，この調査票情報の利用件数が増加している．

一方，匿名データは，調査票情報の直接識別子の情報削除，トップコーディングやリコーディングなどの技法により匿名化されたデータであるが，実社会の傾向を分析するための数理モデルの作成には大いに役立つデータとなっている．著者は，国内外の研究者との共同研究を通して多くの知見を共有することにより，匿名データが研究や教育の品質向上に寄与すると考えている．さらに，大学院生を対象とした「演習（ゼミナール）」において，匿名データを利用することは，データとエビデンスに基づいた実証分析の指導に役立ち，最終目的の修士論文の作成に至るまでの過程において，当該学生が実社会において通用する分析手法を学習するためのよい機会であり，有用性が高くなっている．しかしながら，この有用性に反して，調査票情報の利用者増加による影響か，この匿名データの利用件数が増加していないようである．また，新たな需要が見込まれた「高等教育目的」の利用が，極端に少ないことの影響もあるようである．

さらに，数理・データサイエンス教育強化拠点コンソーシアム（2020）では，「数理・データサイエンス・AI のリテラシーレベルの教育には「学びの動機付け」が重要であり，身近な活用事例や社会の実データ・実課題を用いた演習やグループワークなどを授業に積極的に取り入れることが効果的と考えられる．」（p20）と「講義・演習等による授業上の工夫」で定義している[1]．このことから，大学の演習やグループワークにおいて，公的統計の二次的利用制度を活用した実データを利用することの重要性が述べられている．

本稿では，これまでの「学術研究目的」による匿名データの利用の経験を活かした「高等教育目的」への利用拡大を図るための取組みの準備とその講義方法を提案する．この取組みでは，立正大学データサイエンス学部生を対象に，公的統計の匿名データを利用したビッグデータの実証分析を可能とする PBL（Project Based Learning，課題解決型学習）形式による R 言語を用いた演習講義を実施する．

これまでの匿名データの利用では，小人数の学生を対象とした演習（ゼミナール）は存在するが，10 人以上の受講者を想定した PBL 形式の講義はほとんど存在しない．

なお，匿名データの利用には，利用申請やデータ分析前処理に多くの時間を費やすため，こ

の演習講義においては改正統計法以後の制度を活用し，事前に教授者側で申請を行い，分析前処理においても教授者が実施し，講義時間内での申請や分析前処理は省略する．ただし，この講義の受講者が，より高度な分析を対象とした演習（ゼミナール）を受講する場合，匿名データの利用に関するすべての行程の学習を可能とする．

この取組みの成果として，履修者の履修目標の達成に加え，匿名データの利用申請や講義のための準備作業の負担の軽減，この講義でのデータ分析結果を講義事例として公表することで，これまで匿名データの利用を控えていた教育者や研究者に対する利用促進を可能とする．さらに，匿名データの利用の拡大が，新たなサービス提供へと発展し，公的統計の二次的利用における利用への利便性向上に期待する．

教育機関での匿名データ利用目的については，主に「学術研究目的」と「高等教育目的」があるが，2009 年からの匿名データの利用実績をみると，高等教育目的での利用が進んでいない．2019 年 5 月に，これまでの匿名データの「高等教育目的」の利用者の範囲，申請単位や方法に関して変更されたことを踏まえ，本稿では，今回の変更点や今後の統計教育等の発展の可能性について検討する．

第 2 節の利用可能な匿名データとそれを利用した先行事例では，現在利用可能な匿名データとそれらを利用した講義の事例を述べている．第 3 節では，「演習講義」における「学術研究目的」での匿名データの利用として，一橋大学経済研究所の経済学研究科の大学院生に対する指導について述べ，第 4 節では著者のこれまでの経験を踏まえ，「高等教育目的」における学部生への講義を目的とした匿名データの利用に関する講義内容とその講義に利用するテキストについて述べている．さらに，第 5 節では，本取組みでの期待される効果，第 6 節のまとめ，最後に今後の課題を述べる．

2. 利用可能な匿名データとそれを利用した先行事例

2.1 利用可能な匿名データ

匿名データの利用目的は，以下のいずれかに該当する必要がある．

- 学術研究の発展に資すると認められる統計の作成等（学術研究目的）
- 教育の発展に資すると認められる統計の作成等（教育目的）
- 国際社会における我が国の利益の増進及び国際経済社会の健全な発展に資すると認められる統計の作成等（国際比較統計利活用事業目的）
- デジタル社会形成基本法（令和 3 年法律第 35 号）第 37 条第 2 項第 13 号に規定する特定公共分野に関する統計の作成等であって，国民経済の健全な発展又は国民生活の向上に寄与すると認められる統計の作成等（デジタル社会形成統計利活用事業目的）

引用 https://www.e-stat.go.jp/microdata/anonymity

また，匿名データが利用可能な調査は，以下の 7 調査である．

①総務省　6 調査

国勢調査，住宅・土地統計調査，就業構造基本調査，社会生活基本調査，労働力調査，全国家計構造調査（旧全国消費実態調査）

②厚生労働省　1 調査

国民生活基礎調査

以上，2 省の 7 調査である．なお，これまでの著者の研究では，相対的貧困，少子高齢化，家事労働時間の男女間格差などをテーマにしていたため，社会生活基本調査，国勢調査，全国家計構造調査（旧全国消費実態調査）及び就業構造基本調査は多用してきた調査といえる．

2.2　先行事例

2009 年以降の目的別匿名データの利用件数は，表 1 のとおりである．

また，「ミクロデータ利用ポータルサイト」には，2019 年 5 月の改正統計法施行以降の公的統計の匿名データを利用した申請の内容が公開されている．2019 年 5 月以降の匿名データ利用申請件数は 66 件であり，そのうち，「高等教育目的」による匿名データを利用した講義事例は 5 件である（表 2）[2]．

なお，公開されている情報からは，「学術研究目的」か「高等教育目的」かの区別はできないため，「匿名データの利用目的」欄に講義名が記されている申請を抽出した．しかしながら，大学院生以外の研究者が参画している共同研究プロジェクトなどの場合，学会等の発表を修士論文などに加えて実施する場合があり，その場合においては「高等教育目的」での申請とは限らない点に留意する必要がある．今回の 5 件のうち，1 件が 40 名（教授者 2 名）のグループであり，「国民生活基礎調査」を利用していた．

また，それ以前の 2009 年 4 月から 2019 年 4 月の間において，統計センターの利用実績として公表されている匿名データ利用件数の研究分析の例では，55 件のうち 3 件であった．なお，

表 1　匿名データ利用件数（学術研究目的，高等教育目的）

年	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	合計
合計	23	42	38	35	47	44	47	45	56	63	28	468
学術研究	19	40	33	31	40	43	41	41	50	61	25	424
高等教育	4	2	5	4	7	1	6	4	6	2	3	44

出所　https://www.soumu.go.jp/toukei_toukatsu/index/seido/shoukoku.htm

※年次別の統計法の施行状況報告より、匿名データ利用件数（合計、教育、学術研究別の数）を抽出

表2　匿名データを利用した講義事例

管理番号	提供年月	調査名	提供を受けた人数
40045020200007	2020年11月	国民生活基礎調査	4名
40020020200001	2020年5月	国勢調査 住宅・土地統計調査 全国家計構造調査 （旧全国消費実態調査）	2名
40045020190006	2019年12月	国民生活基礎調査	40名
40020020190013	2019年11月	全国家計構造調査 （旧全国消費実態調査）	2名
40045020190003	2019年11月	国民生活基礎調査	5名

※著者らの利用申請した学術研究目的4件は含まれていない。

この3件は，同一の講義であり，受講者が2〜4名の大学院生である[3].

以上のことから，改正統計法施行以降も，「高等教育目的」による匿名データ利用申請が少なく，さらに受講者が10名以上の講義は1件のみであるため，多くは演習（ゼミナール）などの個別に指導ができる利用方法であり，10名以上の多人数の受講者を対象とするようなデータサイエンティストを育成するためのPBL形式の演習講義が求められている．なお，利用目的別に申請書が用意されているため，利用目的に即した申請書を選択し，その様式に基づいた事項を記載する必要がある[4][5].

・「高等教育目的」の申請事項（申請書からの一部抜粋）

①利用する学校，研究科・学部学科及び授業科目の名称

②授業科目の目的，授業科目で匿名データを利用する必要性及び利用する手法

③授業科目の内容及び匿名データを利用して作成する統計等の内容

④授業科目の開講期間

3.　匿名データを利用した演習（ゼミナール）の概要

3.1　演習開始の経緯

一橋大学経済学研究科における大学院生の演習（ゼミナール）は，修士課程1年次から2年間と2年次からの1年間の履修に分かれるが，多くの学生は2年次から履修することが多い．

そのため，1 年次に履修した科目によって，2 年次の演習の指導教員を選考することが多々あり，著者の演習履修生は「統計調査論」の講義の履修生か，または，演習履修者からの紹介が多くなっている．さらに，著者の演習履修者の多くは，実証分析に関心を持っているため，匿名データを利用した研究に適している．

その他，著者が所属している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では，「政府統計匿名データ利用促進プログラム」を実施しているため，このプログラムを活用し，匿名データを利用することで，学会の参加費及びデータ利用の経費の支援を受けられることも，学生への負担軽減につながり，匿名データを利用した演習を可能としている．

3.2　演習状況

一橋大学経済学研究科の大学院生を対象に，公的統計の匿名データを利用した演習（ゼミナール）を 2017 年以降実施している．毎年度，3 名程度の受講者であるが，各々が研究テーマを選択するため，それらのテーマに関連した調査関連資料等を教材に用いることで，個々のレベルに合わせた指導を可能とした．なお，匿名データの利用には，申請時に許可された利用場所に限定されるため，必要に応じて，講義時間以外にも指導を行った．

2020 年度は，コロナ禍における演習講義の実施であったため，対面での指導の回数が減少したが，オンラインによる指導を増加させることで，継続的な指導が可能となった．

この演習の成果は，最終目標である修士論文であるが，その作成過程において，毎年 9 月に開催される「統計関連学会連合大会」と 11 月に開催する「匿名データ等の利用推進ワークショップ」（一橋大学経済研究所と神戸大学との共催）等での発表を目途とし，修士論文の品質を担保している[6][7][8][9][10][11][12]．なお，このような演習であることから，受講者に「学術研究目的」として匿名データの利用申請を経験させることに加えて，匿名データを厳格に取り扱う意義なども指導した．

3.3　匿名データを講義に利用するメリットとデメリット

匿名データの利用に関するメリットは，以下のとおりである．

①エビデンスに基づく実証分析
②実社会での実践的なデータ分析に対応可能
③ e-stat の公的統計の集計結果よりも詳細な分析が可能
④「高等教育目的」での申請により，講義のための利便性が向上
⑤「高等教育目的」では，講義に限らず，準備における講義用資料作成時にも利用が可能

一方，デメリットは，以下のとおりである．

①最新の調査年次のデータがない調査がある
②分析に使う符号表の見方の理解に時間を要する
③ CSV 形式のデータの分析前処理に時間が掛かる
④利用場所の制限

以上，匿名データを講義で使うためには，これらのデメリットのうち，②・③については，大学などの教育機関において解決する課題である．従って，これらを解決するための方法を以降の章で述べることにする．

4. 高等教育目的での匿名データの利用

4.1 高等教育目的利用に向けて

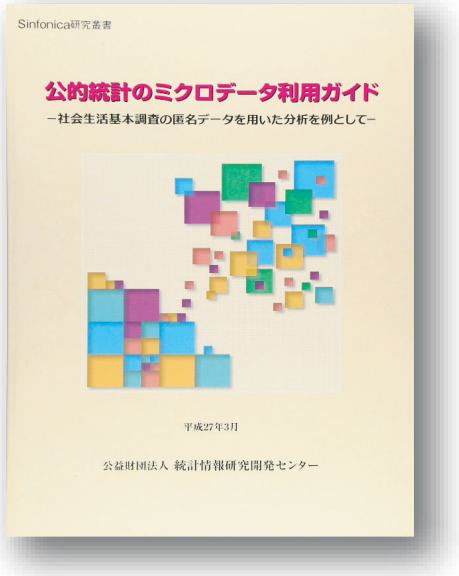
「学術研究目的」として匿名データを利用するには，分析対象調査の選定，利用場所の確保，学生等であれば指導教員の推薦に加え，データの利用料等を準備する必要がある．

毎年度，ゼミナールのような個別の演習の講義開講後，履修者ごとの研究テーマを決定し，その後に匿名データの利用申請することは，実質的な研究に費やす時間が不足する．そのため，受講者ごとに，匿名データの利用申請や符号表などにおける関係書類の見方を指導する講義を別途開設することにより，データ形式の変換や分析に掛かる研究時間の確保を可能とした．

「高等教育目的」での利用では，担当講師が講義のシラバス等に基づき匿名データの利用申請し，当該講義の開講時に履修者全員の誓約書を提出することで，履修者全員がデータの利用ができることになり，講義の内容・目的等が変わらない限り，「学術研究目的」のように申請の段階では利用者全員が明示的でなくても申請することができる．また，コンピュータルームなどの演習講義室等の利用ができることも，「学術研究目的」での利用より，利便性が高いと言える．それゆえ，匿名データの利用目的に応じた申請が，利用者にとっての最大の効用となる．

講義は，PBL 形式として，多くの学生を対象にした匿名データの利用が想定できる．なお，講義では，自主的な学習を期待する場合，「公的統計の匿名データ利用に関するテキスト」を選定し，その内容に基づいてシラバスを作成することが必要であるため，以下にテキストの一例を紹介する．今回の講義では，書籍名『公的統計のミクロデータ利用ガイド―社会生活基本調査の匿名データを用いた分析を例として―』（2015 年 3 月，（公財）統計情報研究開発センター）の利用を想定している（図 1）．この利用ガイドは，公的統計の二次利用制度，申請の方法及び匿名データの利用方法を詳細に記載しており，初見の方々でも，理解しやすい内容となっている．さらに，社会生活基本調査の匿名データを利用した実践的な分析にも取り組んでおり，匿名データの特徴，関連資料の見方及びデータ形式の説明に加え，いくつかの統計分析ソフトウェアによるファイルの読み込みが図入りで解説されているため，誰でも簡単に分析が開始できる．また，このデータを用いた集計結果が正確に集計されているかの検証として，e-Stat の利用についても解説されていることからも，的確な分析ができる．本書での分析事例は，以下の表 3 のとおりである．

これらの分析事例では，分析に用いた集計表の見方，用語の解説に加え，匿名データの符号



参考に，このガイドの目次は，以下のとおりである．
導入部　「二次利用って」どうするの？
第 1 節　二次利用制度の解説及び手続きの進め方
第 2 部　匿名データの使い方
第 3 部　その他の二次利用

図 1　テキスト『公的統計のミクロデータ利用ガイド―社会生活基本調査の匿名データを用いた分析を例として―』（2015 年 3 月，（公財）統計情報研究開発センター）

表 3　分析事例一覧

	テーマ
1	高齢者の年齢 3 階級別行動分析
2	種目ごとの平均行動日数の比較
3	家事・育児に費やす時間の分析
4	趣味・娯楽への生活時間配分の分析　その 1
5	趣味・娯楽への生活時間配分の分析　その 2

表の見方や集計及び結果の解釈が掲載されている．なお，結果の解釈では，集計表及び図を用いた解説のため，回帰分析等の分析事例はないが，公的統計の二次的利用の初学者が匿名データを利用することに関しては適した内容である．

ただし，2015 年 3 月の刊行後，改訂が行われておらず，改正統計法が施行された 2019 年 5 月以降の状況が反映されていない．そのため，現在，（公財）統計情報研究開発センターにおいて改訂版を準備中である．

その他，必要に応じて，各統計調査の特徴や符号表等の関連資料をわかりやすくまとめるこ

表4　匿名データを利用した演習のシラバス（抜粋）

	内　　容
科目名	統計学実習 IV（３年次）
授業の目的 （抜粋）	主に公的統計のいくつかを題材に、公的統計へのビッグデータ活用が経済・社会・意思決定等に対して有効な統計を提供するために大きな役割を果たすことの実際を具体的な演習を通して学ぶ
到達目標	公的統計の匿名データ等のビッグデータにおいて、解析ソフトを用いてデータの集計前処理、集計及び分析を実施し、分析レポートの作成ができる。

とや講義の履修者のレベルに合わせて，①分析のためのデータ前処理，②公表済み結果表と同一表の作成，③オリジナル集計表の作成，④回帰分析，⑤時系列分析，⑥機械学習，等に特化したシラバスにすることにより，幅の広い講義が可能となる．

4.3　講義内容等

匿名データを「高等教育目的」に利用する講義は，立正大学データサイエンス学部の統計学実習 IV（3 年次）である．本講義の達成目標では，「公的統計の匿名データ等のビッグデータにおいて，分析ソフトを用いてデータの集計前処理，集計及び分析を実施し，分析レポートの作成ができる．」としており，実社会に即した分析を可能とするための目標を掲げている（表4）．

利用する匿名データは，総務省の「社会生活基本調査」と「国勢調査」を想定している．

講義方法等は，以下のとおりである．

・履修者全員からの匿名データの利用のための誓約書を受領
・コンピュータルームのみでの利用
　※コロナ禍であるため，収容人員は最大 40 名程度
・PBL 形式による小グループ制（利用責任者を任命）
　※ SA（スチューデント・アシスタント）の人数に応じて，募集人員は制限する．
・講義内のみでの利用時間制限
・R 言語を用いたデータ分析
・講義終了後のアクセス権や元データの削除
・その他，総務省が定める匿名データ利用のルールを順守，適正管理措置に関する指導

4.4　分析手順

第一に，研究テーマに基づいた背景と目的に基づき，研究に利用する統計調査を選定する．

その後，以下の手順で分析する．

①統計調査に関連する調査方法，集計結果及びトピック等の情報を取得

②「政府統計の総合窓口（e-Stat）」から詳細な集計結果情報を取得

③匿名データを利用した分析

④各自の研究テーマに即した詳細な結果表やモデル式による分析

なお，必要に応じ，匿名データ以外のデータも加え，設定したテーマに応じた詳細な分析を試行する．

4.5　講義のための環境とその利用制限

立正大学データサイエンス学部では，演習講義のためのコンピュータルームが設置されている．この演習ルームにおいてはR言語によるデータ分析が可能な環境をそろえている．履修者は，シンクライアント形式のシステムを利用することで，情報漏洩することなく，安全に演習講義を実施することが可能となる．また，履修者は，講義時間以外の利用ができないように制限する．

4.6　データ分析のための前処理（事前準備で対応）

上記4.1で紹介したテキストに基づき，データ分析の前処理の一例として，社会生活基本調査であれば，以下のとおりとなる．なお，この前処理は，教授者が講義前に実施することを想定している．

①符号表の「項目番号」とデータの列数との照合

・すべての列とデータ内容を符号表と対応させる．

・同調査は「繰り返し」と限定項目（特定の条件を満たす人のみ回答）が非常に多いため，すべての列とデータ内容を符号表と対応が困難である．

②使用したい列について，列名をつけ，カテゴリカル変数ならば各水準にラベルをつけ（＊），欠測値をつけ，融合すべきものは融合する（＊＊）．

・（＊）たとえば，「01」，「02」，…という水準があるが，これがなにを表す連番かわかないため，符号表などを参考にして，「01」→「義務教育」，「02」「03」→「高等教育」…などと，便宜的にラベルをつける．

・（＊＊）たとえば，「有業か無業か」という列の他に「（うち有業のみ）有給休暇の日数」という列がある場合，1つの列に融合して「有給休暇の日数」（0は無業の意味）とする．ただし，融合に際して回答の解釈に誤りがある場合があるため，何をどのような解釈で融合したのかを確認する．

以上の作業により，R言語等で分析可能なデータ構造を編成する．

4.7　R言語による分析事例

演習講義で使用するテキストにある分析事例（表3）を，R言語で作図をする．

上記，図2から図5は，講義において解説する分析事例の一部であるが，これ以外の表3の

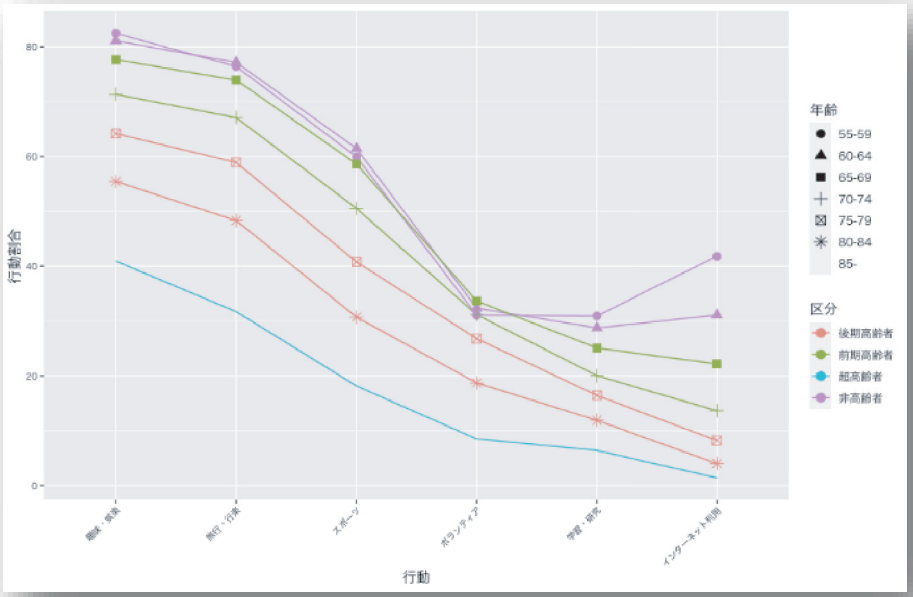


図２　高齢者の自由時間における余暇の状況①

```r
####################
# p.14 （図5）年齢階級別行動者率
####################
NENREI <- c("55-59","60-64","65-69","70-74","75-79","80-84","85-")
bigdat %>% select(年齢,matches("総_.*有無")) %>%
  filter(年齢 %in% NENREI) %>%
  pivot_longer(cols = -年齢, names_to = "行動", values_to = "行動割合") %>%
  mutate(
    across(行動, ~str_replace_all(., "趣味娯楽_総_活動有無", "趣味・娯楽") ),
    across(行動, ~str_replace_all(., "旅行_総_活動有無", "旅行・行楽") ),
    across(行動, ~str_replace_all(., "スポーツ_総_活動有無", "スポーツ") ),
    across(行動, ~str_replace_all(., "ボランティア_総_活動有無", "ボランティア") ),
    across(行動, ~str_replace_all(., "学研_総_活動有無", "学習・研究") ),
    across(行動, ~str_replace_all(., "ネット_総_利用有無", "インターネット利用") ),
    across(行動, ~factor(.,
      levels = c("趣味・娯楽","旅行・行楽","スポーツ","ボランティア","学習・研究","インターネット利用") ))
    ) %>%
  group_by(年齢,行動) %>%
  summarise(across(行動割合, ~{ (sum(.=="有")/length(.))*100 }), .groups = "keep") %>%
  mutate(
    区分= case_when(
      年齢 %in% c("55-59","60-64") ~ "非高齢者",
      年齢 %in% c("65-69","70-74") ~ "前期高齢者",
      年齢 %in% c("75-79","80-84") ~ "後期高齢者", TRUE~"超高齢者" )
    ) %>%
  ggplot(aes(x=行動, y=行動割合, group=年齢, shape=年齢, color=区分)) +
  geom_line() + geom_point(size=3)
```

図３　高齢者の自由時間における余暇の状況①（R コード）

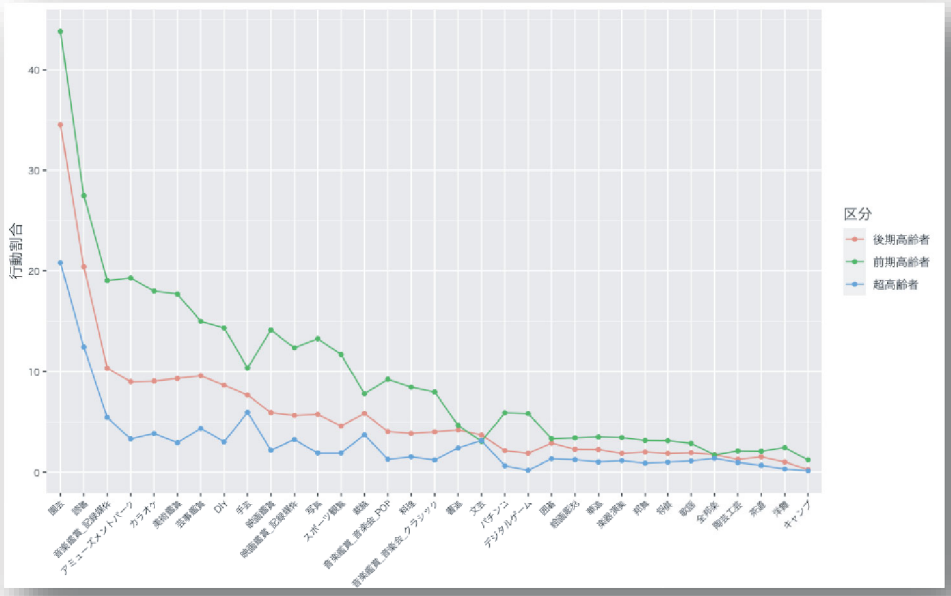


図４　高齢者の自由時間における余暇の状況②

```
#####################
# p.15 （図6）趣味・娯楽の種目別行動率
#####################
NENREI <- c("65-69","70-74","75-79","80-84","85-")
bigdat %>% select(年齢,matches("趣味娯楽_.*頻度")) %>%
  filter(年齢 %in% NENREI) %>%
  pivot_longer(cols = -年齢, names_to = "行動", values_to = "行動割合") %>%
  drop_na %>%
  mutate(
    across(行動, ~str_remove_all(., "趣味娯楽_|_頻度") ),
    区分= case_when(
      年齢 %in% c("65-69","70-74") ~ "前期高齢者",
      年齢 %in% c("75-79","80-84") ~ "後期高齢者", TRUE~"超高齢者" )
  ) %>%
  group_by(区分,行動) %>%
  summarise(across(行動割合, ~{ (sum(.!="0")/length(.))*100 }), .groups = "keep") %>%
  ggplot(aes(x=reorder(行動,-行動割合), y=行動割合, group=区分, color=区分)) +
  geom_line() + geom_point(size=2, shape=20)
```

図５　高齢者の自由時間における余暇の状況②（R コード）

分析を各グループで実施し，その成果を最終講義において発表する．

5. 期待される効果

5.1 実証分析の経験

講義において，約 120 万レコード（国勢調査）や約 880 変数（社会生活基本調査）もあるデータを利用した分析ができることと，それらの分析経験は，データサイエンスに関する知識習得やデータ分析手法の選択肢を拡張することが可能となる．

5.2 高度な分析手法の習得

本稿の取組みにおいて，匿名データの利用のための最低限の知識を習得することにより，さらに高度な実証分析が可能となり，各学会等が主催する研究会での発表や卒業論文に適用するための演習講義がより効果的に実施可能となる．1 講義は半期（15 回）の短期間であるため，3 年次から卒業までの 2 年間の複数講義での匿名データを利用することで，データ分析やその結果の説明ができるデータサイエンスのための人材育成に期待ができる．

5.3 利用データの汎用化

データ分析前処理として，講義に利用するデータの種類（オリジナル，記号等を処理した CSV データ，R のオブジェクトとしてのデータ，変数を絞り込んだデータ，等）により，データ分析の難易度の調整を可能し，講義時間内に課題がクリア出来るようにする．

この成果を，テキスト，またはシステム化を行い公開して，多くの利用者が利用できるようにすることで，匿名データの高等教育目的による利用申請件数の増加が期待できる．

6. まとめ

公的統計の匿名データの利用において，筆者のこれまでの「学術研究目的」を活かした上で，データサイエンティストを育成するため，PBL 形式の多人数での講義を対象とした「高等教育目的」での利用のための講義事例について述べた．

本稿では，匿名データを講義に利用するために達成目標や演習する内容を絞り込む必要があること，加えて匿名データの利用に適したテキストが必要であることについて述べた．本講義は，ビッグデータや AI などに関心がある学生等の人材育成に役立つことを想定しており，このような講義が行われることにより，今後の公的統計の二次的利用の利便性の向上や，日本の良質なビッグデータ分析の推進に寄与すると期待される．

7. 今後の課題

継続的な講義を実施するため，以下の項目について検討する．

検討事項は，①現存する 7 調査に対応するテキスト等の作成，②データ分析のための前処理に手間が掛かるため，一橋大学経済研究所が提供する「統計データ変換ツール」[13]など，Stata

やRなどのデータ形式の変換ツールの活用の検討，③匿名データの理解を深める取組みとして，符号表の見方（簡単な語句の説明シートの用意）など，理解しやすい資料の作成である．

その他，匿名データの格納項目について，①融合（上位の項目の値によって下位の項目の値が決定する）すべき列は，融合後の値とするなど，加工した項目の格納，②FILLER，集計用乗率，及び値が1つしかない列など，集計用の詳細情報の明確化，③「未記入」と「不詳」と「マルチマーク」と「限定項目の回答対象外」の判別の厳密化，④各項目における欠測値の定義の明確化である．

以上，今後，準備・適用可能な事項より検討し，改善する予定である．

謝辞

今回の演習講義を実施するにあたり，『公的統計のミクロデータ利用ガイド―社会生活基本調査の匿名データを用いた分析を例として―』（発行：公益財団法人統計情報研究開発センター（Sinfonica））の改訂版の作業に携わって頂きました公益財団法人統計情報研究開発センターの編集者の方々に感謝致します．

参　考　文　献

［1］「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム～データ思考の涵養～」
2020年4月，数理・データサイエンス教育強化拠点コンソーシアム
http://www.mi.u-tokyo.ac.jp/consortium/pdf/model_literacy.pdf
［2］ミクロデータ利用ポータルサイト
匿名データの利用実績　提供状況一覧（匿名データ）
https://www.e-stat.go.jp/microdata/jisseki/anonymity? page＝1
［3］利用実績（改正統計法施行前）
https://www.nstac.go.jp/services/jisseki.html
［4］申請書　学術研究目的関係（Excel：154KB）
https://www.nstac.go.jp/services/2ji/tokumei_r-youshiki01-1-1.xlsx
［5］申請書　教育目的関係（Excel：155KB）
https://www.nstac.go.jp/services/2ji/tokumei_r-youshiki01-2-1.xlsx
統計関連学会連合大会におけるゼミ生の発表
［6］張加斌，白川清美「AIによる職業淘汰とワークライフバランスへの影響」2020年度統計関連学会連合大会（オンライン開催），2020年9月
［7］黎雨西，白川清美「日本の就労率と管理職率におけるジェンダーギャップ」2020年度統計関連学会連合大会（オンライン開催），2020年9月
［8］高宇，白川清美「日本の少子高齢化における外国人労働者数の動向と今後の展開」2020年度統計関連学会連合大会（オンライン開催），2020年9月
［9］朱永楽，白川清美「東京都内における空家率と商業用住宅価格の相関性分析」2019年度統計関連学会連合大会（於：滋賀大学）2019年9月
［10］張一鼎，白川清美「訪日観光客がもたらす地域観光業における労働市場への影響」2018年度統計関連学会連合大会（於：中央大学）2018年9月
［11］李敏杰，白川清美「家電製品が近未来における日本の女性に与える家事労働時間への影響」2018年度統計関連学会連合大会（於：中央大学）2018年9月

[12] 平川竜成，白川清美「匿名データに基づいた生活行動における嗜好が未婚率に与える影響の分析」2017年度統計関連学会連合大会（於：南山大学）2017年9月　　分析に役立つツールの提供
[13] 一橋大学経済研究所が提供する「統計データ変換ツール」
https://rcisss.ier.hit-u.ac.jp/Japanese/micro/study03.html