様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ４ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０３６６５

研究課題名（和文）統計調査における外れ値の検出とそれへの対応に関する理論的・実証的研究

研究課題名（英文）Detection and handling of outliers in surveys

研究代表者
西郷　浩（Saigo, Hiroshi）

早稲田大学・政治経済学術院・教授

研究者番号：００２０５６２６

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：　統計調査データを集計する際に発生する外れ値の検出と処理について研究した。実際の調査において発生する外れ値の性質を調べるため、経済統計研究会を定期的に開催した。1年に4回、合計で12回開催した。外れ値発生の頻度が高く、かつ、外れ値処理の影響が大きい統計を中心に、統計作成者を講師としてまねき、討論した。経済統計研究会における討論にもとづいて論文を構想した。その成果は、おもに『統計』（日本統計協会）に発表したほか、現在、2つの論文を学術誌に投稿している。

研究成果の学術的意義や社会的意義
統計調査における外れ値は、集計結果に影響を及ぼす。その的確な処理は、調査結果の集計精度向上に役立つ。統計調査における外れ値は、調査対象（世帯調査・事業所調査）に応じて性質が異なる。また、外れ値の発生頻度は、統計基準（産業分類など）によっても左右される。数理的な性質とともに、統計調査の周辺の条件も考慮しながら外れ値の検出と処理方法を研究することが本研究の特徴である。この研究の結果、統計調査の集計結果の安定することが期待できる。

研究成果の概要（英文）：We focused on methods for detecting and handling outliers in survey statistics. To obtain practical information on outliers in real survey data, we organized a meeting called Economic Statistics Meeting on a regular basis, four meetings in a year, twelve in three years. In the meeting, we discussed survey statistics in which outliers occur frequently and influence estimation tremendously through questions and answers with those who make the statistics. These discussions are a source of our research. Some papers are presented in Statistics published by the Japan Statistical Association. Currently, two manuscripts are under review in refereed journals.

研究分野：経済統計

キーワード：統計調査法　外れ値　世帯調査　事業所調査

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究の目的は、統計調査において発生する外れ値の検出とそれへの対応について、統計理論に基づいて統一的な基盤を与えることにある。外れ値は、あらゆる統計調査において発生する。事業所の調査において、業種・規模が似通った事象所の中に、販売額が他より極端に大きい少数の事業所が存在することは稀でない。世帯の調査においても、世帯人員の数や構成が類似した世帯の中に、支出額が極端に大きい世帯が存在することはむしろ普通である。外れ値の発生理由は多様である。桁数の誤解による回答者の記入ミスのように、回答そのものに誤差が含まれている場合がある。また、購入頻度の少ない高額耐久消費財を購入した世帯の支出額のように、回答に誤差が含まれていなくても、他の観察値と極端に異なる値が観察される場合もある。

どのような理由であれ、少数の外れ値が、推定値、とりわけ母集団合計の推定値に大きく影響を及ぼす場合が少なくない。外れ値の影響によって推定値が大きく変動する結果、推定値の時系列的な変動が非現実的なほど大きくなることもある。記入ミスのように、修正を是とすることが自明の場合ばかりでなく、記入そのものに誤差がないとしても、推定値の変化が極端に大きいときには、推定の段階で外れ値に対応する必要が生じる。

２．研究の目的

統計調査における外れ値処理には、固有の困難がある。その理由は次の 2 点に要約される。ひとつは、(a) 世帯調査においても事業所・企業調査においても、観察値の分布の右裾が長い（右に歪んでいる）ことが多いこと、もうひとつは、(b) 母集団の合計が加算で計算されていること、である。

まず、(a) 強い右への歪みが困難を引き起こすことの理由について説明する。強い右への歪みは、外れ値の検出とそれへの対応との両方で理論的な難問となる。外れ値の検出の標準的な手法では、何らかの理論分布を想定して、その理論分布から導かれた臨界値よりも大きな観察値を外れ値とみなす。しかし、分布の右裾の観察値の個数が少ない場合には、理論分布を適切に選ぶことが難しい。このため、選択した理論分布によって、臨界値が異なるため、外れ値の安定的な検出規則を決定することが難しくなる。さらに、適切に理論分布を特定して外れ値が検出できたとしても、推定の段階で、外れ値をどのように処理するかが困難である。外れ値への標準的な対処方法としては、頑健推定が挙げられる。その典型的な方法は、一定の基準にしたがって外れ値のウェイトを減少させ、推定値への影響を削減することである。しかし、これまでの頑健推定の研究は、対称な分布における外れ値への対処方法を中心に進められており、右への歪みが強い分布に適した頑健推定の研究の余地が多く残されている。

もうひとつの困難の理由は、(b) 母集団の合計が加算によって計算されることである。このことは合計の定義から当然である。しかし、このことは理論上の困難となる。たとえば、右への歪みが強い分布への古典的な対処方法は、変数を対数変換して分布を近似的に対称にすることである。想定される理論分布によっては、対数変換によって理論分布の母数が推定しやすくなることもある。たとえば、対数正規分布を想定できる状況であれば、標本の観察値の対数の平均によって、母平均をより正確に推定できる。ところが、対数変換した観察値の平均は、幾何平均であって、母集団合計を母集団のサイズで除した算術平均とは異なる。幾何平均から母集団合計をどのように推定するかは自明ではない。統計調査における推定対象は、販売額合計のように、もともと加算で定義されていることが多い。推定対象が合計で定義されているという条件のもとで、外れ値に対応しなければならない。

こうした、統計調査における外れ値に固有の問題に対処できる手法を考案することが本研究の目的である。

３．研究の方法

以上のような、統計調査における外れ値に固有の問題について情報を収集するため、統計調査の実施者を講師として招き、個別の統計作成に発生する実際の問題、とりわけ外れ値の処理について報告していただき、それに基づいて討論する形式の研究会、経済統計研究会を定期的に開催した。

経済統計研究会の討論をもとに、論文を作成した。

４．研究成果

(1)経済統計研究会の開催

3 年間で 12 回の経済統計研究会を開催した。

(a)「国民経済計算の平成 23 年基準改定の推計結果等について」（2017 年 6 月 24 日）

(b)「日本の将来推計人口」（2017 年 10 月 28 日）

(c)「平成 28 年社会生活基本調査 概要と結果について」、「我が国の勤務間インターバルの状況：平成 23 年社会生活基本調査の結果から」（2017 年 12 月 2 日）

(d)「米国センサスにおける無回答の処理と公開用ミクロデータ作成を参考とした我が国における国勢調査の補定処理と匿名化に関する提案」（2018 年 2 月 3 日）

(e)「地域別人口の将来見通しと地域差の要因：『日本の地域別将来推計人口（平成 30 年推計）より』」、「今後の世帯数と世帯構成の見通し：全国世帯推計（2018 年推計）の結果から」（2018 年 7 月 21 日）

(f)「創設の想いを受け継ぎ刻む新たな歴史：平成 28 年経済センサス-活動調査を終えて」（2018 年 10 月 13 日）
(g)「財務省法人企業統計調査について」（2018 年 12 月 18 日）
(h)「事業所母集団データベースの概要と母集団情報の精度向上に向けた取り組み」（2019 年 3 月 2 日）
(i)「平成 30 年住宅・土地統計調査の実施状況等について」（2019 年 7 月 6 日）
(j)「日本の生産物分類：現在までの取り組みと今後の課題」（2019 年 9 月 28 日）
(k)「毎月勤労統計の不正と公的統計の改善に向けて：日本統計学会『公的統計に関する臨時委員会報告書』」
(l)「国勢調査ことはじめ」（2020 年 2 月 1 日）

(2)ミクロデータによる社会生活基本調査の特別集計と外れ値処理

研究協力者（仲村敏隆）とともに、総務省統計局のオンサイト施設を利用して、社会生活基本調査のミクロデータを特別集計した。その際、社会生活基本調査において発生する外れ値に適した処理を採用した。その研究成果を、「官民オープンデータ利活用の動向及び人材育成の取組(2019 年度)」（2019 年 11 月 15 日 統計数理研究所）において報告した。

(3)論文作成

経済統計研究会における討論に基づいて、2 つの論文を作成し、現在、査読中である。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 西郷浩	４．巻 16(1)
２．論文標題 文化芸術活動の分析のための社会生活基本調査	５．発行年 2019年
３．雑誌名 文化経済学	６．最初と最後の頁 15-19
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 西郷浩	４．巻 69(7)
２．論文標題 生活時間からみた高齢単身者	５．発行年 2018年
３．雑誌名 統計（日本統計協会）	６．最初と最後の頁 50-53
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 西郷浩	４．巻 69(4)
２．論文標題 生活時間からみた単身者	５．発行年 2018年
３．雑誌名 統計（日本統計協会）	６．最初と最後の頁 20-25
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 西郷浩	４．巻 68(6)
２．論文標題 統計から見た大学進学	５．発行年 2017年
３．雑誌名 統計（日本統計協会）	６．最初と最後の頁 51-54
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 仲村敏隆
２．発表標題 社会生活基本調査を用いたデジタルゲームの需要行動に関するコーホートの視点からの基礎的分析
３．学会等名 官民オープンデータ利活用の動向及び人材育成の取組(2019年度)
４．発表年 2019年

〔図書〕　計1件

１．著者名 日本統計協会	４．発行年 2017年
２．出版社 日本統計協会	５．総ページ数 366（うち、第12章余暇活動 pp. 122-131 を西郷が担当）
３．書名 統計でみる日本2018	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	仲村　敏隆 (Nakamura Toshitaka) (32689)	早稲田大学・大学院経済学研究科・博士課程学生	研究計画書に記載したミクロデータの集計・分析を担当