国民生活基礎調査

西郷　浩
（早稲田大学教授）

1　はじめに

今回，労働統計という観点から厚生労働省の『国民生活基礎調査』について執筆する機会をいただいた。しかし，私は労働経済学の専門家ではない。そこで，労働統計としての分析というよりは，『国民生活基礎調査』を題材として，本格的に個票データを分析する前に済ませるべきと私が考える事柄について述べることにする。

2　個票データを分析する前に

(1) 調査票を見る

個票データは調査票から作成される。そのため，調査票から分かる情報しか個票データには含まれない。したがって，個票データを分析する前に，調査票を確認して個票情報の具体的な内容を正確に把握する必要がある。

たとえば，就業状態について，2019年の『国民生活基礎調査』（大規模調査）では，世帯票の質問13において，「5月中の仕事の状況」として「収入を伴う仕事を少しでもした方は『仕事あり』，まったく仕事をしなかった方は『仕事なし』の中からお答えください」と尋ね，「仕事あり」は4項目（例：主に仕事をしている），「仕事なし」は3項目（例：通学）の中から選択することになっている。つまり，就業状態に関する調査期間を1カ月（5月中）としたアクチュアル方式によって就業状態が把握されている。これに対して，平成16（2004）年の調査（大規模調査）では，世帯票の項目（18）において，「『所得を伴う仕事の有無』と『就業希望の有無と理由』」として，「仕事あり」は4項目（例：主に仕事をしている），「仕事なし」は3項目（例：通学のみ）の中から選択することになっている。2019年調査と比べると，アクチュアル方式であることを明示する説明文がないこと，選択肢の文言が異なる場合があること，調査票のレイアウトが異なること（例：2019年調査では就業希望の有無・理由に関する質問が，質問18として別の個所に誘導されている）などの違いがある。なお，調査票だけからは分からないけれども，従来の他計式（調査員による聞き取り）調査から，平成19（2007）年に自計式（回答者が自分で調査票に記入する）調査に移行している。

調査票における質問の文言や調査票のレイアウトなどは，調査項目の明確化や調査環境の変化などに合わせて常に見直される。時間に余裕があれば，試験調査によって調査票の変更による調査結果への影響を定量的に評価してから調査票が変更される。しかし，それができない場合も多い。このため，調査票の変更が回答にどれほどの影響を及ぼすかを定量的に述べるのは難しいことが多い。しかし，分析者は，調査票における差異を認識し，その影響について思料しておくべきだと私は思う。

(2) 標本設計を確かめる

『国民生活基礎調査』の標本設計は，公的統計の中でも極めて独特と私には見える。そして，複数の調査票が調査の各段階で配り分けられる。このため，現行の調査の設計の下でそのまま実行できる分析と工夫を経て実行できる分析がある。以下では，まず標本設計の概略を述べて，実行できる分析の例を述べる。

『国民生活基礎調査』の集計項目と標本抽出の実行方法，標本サイズなどは大規模調査年と簡易調査年で異なる。大規模調査年においては，5種類の調査票（世帯票，健康票，介護票，所得票，貯蓄票）が配布され，全国および都道府県別の集計に目標精度が定められるように標本が抽出される。簡易調査年においては，2種類の調査票（世帯票，所得票）が配布され，全国の集計に目標精度が定められるように標本が抽出される。大規模調査年においては，(1) 世帯票と健康票が配布される世帯は，層別系統集落抽出法によって抽出される。抽出単位は国勢調査区であり，抽出された国勢調査区内のすべての世帯に調査票が配布され

る。(2) 介護票は，(1) で抽出された国勢調査区から約半分の国勢調査区を系統抽出して，抽出された国勢調査区内の要介護等に該当する世帯に配布される（二相抽出法)。(3) 所得票と貯蓄票は，(1) で抽出された国勢調査区のうち (2) で抽出されなかったものから系統抽出し（ここまでは二相抽出法)，抽出された各国勢調査区からその国勢調査区に含まれる調査単位区を無作為抽出して，抽出された調査単位区内のすべての世帯に配布される（全体として多段抽出法)。特に (3) は，二相抽出法と多段抽出法が組み合わせられている。

次に，複数の調査票の配布状況によって，そのまま実行できる分析と工夫が必要な分析の例を考える。大規模調査年の『国民生活基礎調査』に回答した世帯は，(a) 上記 (1) だけで調査された世帯（世帯票と健康票に回答)，(b) 上記 (2) で調査された世帯（世帯票と健康票，介護票に回答)，(c) 上記 (3) で調査された世帯（世帯票と健康票，所得票，貯蓄票に回答）の 3 種類に分かれる。特に，所得票と介護票の両方を配布される世帯はない。報告者負担を軽減するために，これら 2 つの調査票は同時に配布されない。したがって，個票データには，所得票と介護票の組み合わせに関する情報は含まれていない。

このため，調査の設計を所与として世帯ないし世帯員の経済状況と介護の実態との関連を分析するためには，(i) 世帯票に含まれる経済状況に関連する項目を利用する，(ii) 種々の前提条件の下に擬似的なデータを作成して分析する，などが実際の対応となる。(i) については，調査月の家計支出総額や公的年金の受給状況，教育，就業状態，仕事の内容，などへの回答を利用して，(b) の個票データを分析する。(ii) については，傾向スコアなどを利用して，(b) と (c) を統計的に照合することによって擬似的に所得票と介護票の両方の項目を持つ世帯ないし世帯員を創出して分析する。ただし，統計的照合のための前提条件が成立しなければ分析結果に偏りが生じることを認識しておかなければならない。

(3) 集計表を眺める

『国民生活基礎調査』は，それに基づいて作成される『国民生活基礎統計』（基幹統計の 1 つ）が表章に耐える精度を保つように設計されている。個票データの分析は，『国民生活基礎統計』を出発点として細部に分け入ることに等しい。したがって，個票データの分析を始める前に，集計表を眺めて，自分が分析する分野の周辺の風景を思い描けるようにしておくのがよい。

特に，『国民生活基礎統計』には，国民の生活に資する観点から，世帯の属性などに独自の項目が設けられている。一例として，世帯における乳幼児数と家計支出との関係を取り上げる。

表 1 から，おおよそ以下のことが読み取れる。世帯人員の総数の欄を見ると，乳幼児数の数が多くなると平均家計支出額が大きくなる傾向がある。しかし，世帯人員を一定とする（たとえば，世帯人員を 4 人とする）と，2 つの例外を除いて，乳幼児数が増えると平均家計支出額が小さくなる傾向がある。

世帯人員が一定であるときに乳幼児数と平均家計支出額の相関が負になる理由は，おそらく，乳幼児の数が多いほど世帯主が若く，所得水準が低い世帯が多いためだろう。ミクロデータを用いて，世帯人員だけでなく，世帯主の年齢やその他の所得と関連あるような項目（たとえば，世帯員の就業状態や就業日数）を調整すると，乳幼児数と家計支出との関係が表 1 とは異なる様相を呈するかもしれない。

表 1 において，世帯人員が 3 人で乳幼児数が 2 人の世帯の平均家計支出額が，周辺の範疇のそれに比べて突出している。逆に，世帯人員が 4 人で乳幼児数が 3 人以上の世帯の平均家計支出額は極端に小さい。これらの範疇に属する世帯はもともと少ないと予想できる。結果的に，標本に抽出される世帯の数も少なくなり，平均家計支出額の推定値の散らばりは大きくなる。ミクロデータを分析するときにも，これらの世帯のデータが外れ値として分析結果に影響を及ぼしうることに注意しなければならない。

(4) 他の統計と比べる

『国民生活基礎統計』と他の類似の統計とを比較することは，それらの統計の癖を知るのに有効である。有名な例では，『国民生活基礎統計』から計算した相

表 1　世帯人員・乳幼児数別 1 世帯当たり平均家計支出額

（単位：万円）

乳幼児数	世帯人員						
	総数	1 人	2 人	3 人	4 人	5 人	6 人以上
総数	23.6	15.8	23.6	27.7	29.5	32.1	35.4
0 人	23.3	15.8	23.7	28.2	30.6	33.9	36.9
1 人	25.9	—	17.2	24.4	26.3	27.6	34.2
2 人	27.4	·	—	34.7	25.8	29.9	31.8
3 人以上	26.5	·	·	—	15.2	25.4	29.9

注：1）表中で「・」は観察値がありえないこと，「—」は観察値がなかったことを表す。
2）平均家計支出額は 5 月中のものである。
出所：厚生労働省『国民生活基礎統計』（2019 年調査 世帯表 第 056 表）

対的貧困率が総務省の『全国消費実態統計』（『全国家計構造統計』の前身）から計算したそれよりも６ポイント程度高いことが指摘されている。ただし，水準に差があるものの，変化の方向は両統計で同じになることも観察できる。

似たような統計であるにもかかわらず差が生じる理由は，調査対象や調査時点，調査期間，調査方法，調査項目の定義，調査方法に左右される無回答誤差・回答誤差の発生の仕方，などの違いに求められる。ただし，どの原因がどの程度の差をもたらすかを定量的に測るのは難しい。したがって，それらの違いを認識しつつ，実際に２つの統計を同じ時点で比べ，そこに差があれば複数の時点でさらに比較する。差が安定的であれば，標本誤差以外の何らかの系統的な原因によって相違が発生していると考えられる。

ここでは，『国民生活基礎統計』と総務省の『労働力統計』における，男女・配偶関係・年齢階級別の就業状態を比較する。具体的には，『国民生活基礎統計』（2019 年調査）における男女・配偶者の有無・年齢階級別人口に占める「仕事あり」の者の割合と，『労働力統計』（2019 年５月調査）における男女・配偶関係・年齢階級別人口に占める就業者の割合（就業率）を比べる。『国民生活基礎統計』における配偶関係は配偶者あり・なしの２種類，『労働力統計』のそれは有配偶，未婚，死別・離別の３種類である。『国民生活基礎統計』における「仕事あり」の定義は２（1）で述べた。より詳しくは，一時的に仕事を休んでいる者も「仕事あり」に記録される。『労働力統計』における就業者は，調査週間（月末一週間）に収入を伴う仕事をしていた者（従業者）と職を持ちながらも仕事を休んでいた者（休業者）から成る。調査方法の違いの他に，就業状態を調べる期間の長さが両調査で異なる。

図１から，２つの統計における男女・配偶関係・年齢階級別の就業状態はきわめて似ている。この項目については，調査方法の違い等が顕在化していないように見える。

３　個票を分析するときに

（1）集計コードを確認する

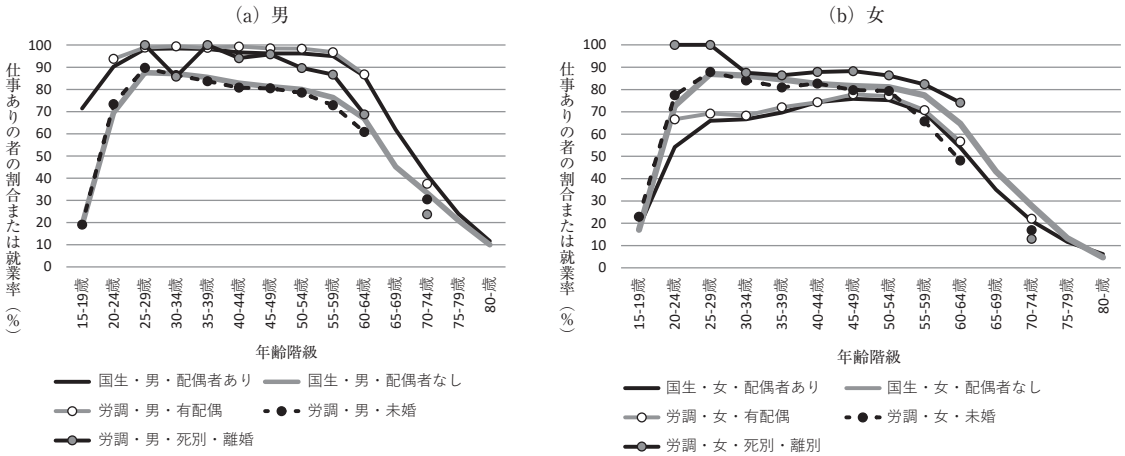
かなり以前に個票データを初めて自分で分析したとき，渡されたコード表に記載されていない複数の記号がデータの中に含まれていて驚いたことがあった。最近では，ミクロデータとしての公的統計の利用環境の整備が進んでいる。したがって，私の経験はもはや昔話なのかもしれない。

しかし，分析の初期段階で利用するデータを確認することは今でも必要だと私は思う。データが正常に読み込め，その中に登場するすべての記号を把握する。公的統計から作成されるミクロデータは，規模が大きいながら定型である。その点は，非定型のビッグデータと異なる。分析の出発点ですべての記号を自分で掌握していれば，安心して分析を始められる。

（2）ヒストグラムを作成する

本格的な分析の前に，量的変数と質的変数のいずれについても，変数の分布を確かめる。そのために，量的変数についてはヒストグラムを作成する。要約統計量だけでなく，変数の分布を見ることで分析のヒントが得られることがある。また，ヒストグラムによって，データの品質を確かめて，分析の際に注意すべき

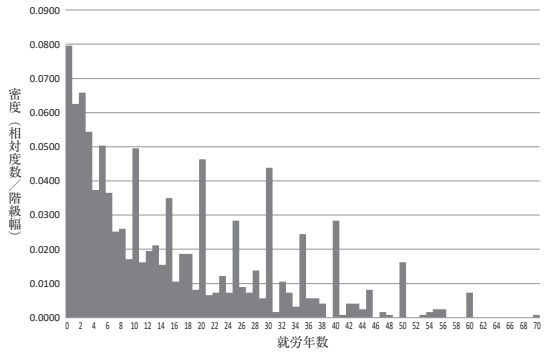
図１　男女・配偶関係・年齢階級別「仕事あり」の者の割合と就業率



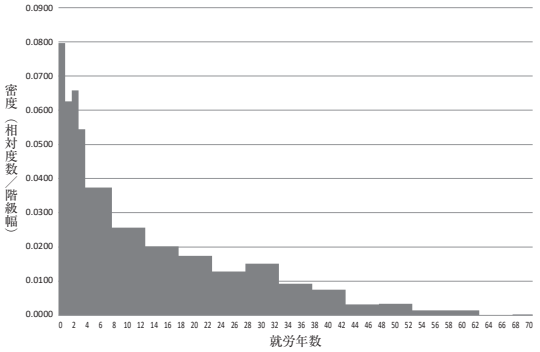
注：1）標本の小さいために結果が不安定であると思われる場合を除いてある。
　　2）『労働力統計』における年齢階級 65 歳以上の就業率を，年齢階級 70-74 歳に表示する。
出所：厚生労働省『国民生活基礎統計』（2019 年調査），総務省『労働力統計』（2019 年５月調査）

図２　JGSS 2005 における就労年数のヒストグラム

（a）原データ（階級幅１のヒストグラム）



（b）就労年数４年以上を階級化したヒストグラム



出所：日本版 General Social Surveys〈JGSS-2005〉

点が明らかになることもある。

『国民生活基礎統計』ではないけれども，一例として，JGSS2005（謝辞参照）における就労年数のヒストグラムを図２に示す。

図 2(a) から，就労年数が４年以上になると，末尾が０と５の就労年数の出現頻度が周辺と比べて突出することが分かる。記憶に基づく回答では，このような現象が生じやすい。言い換えれば，就労年数が長いと回答誤差が大きくなる。この場合，たとえば，就労年数の１年の変化が賃金率にどのくらいの変化をもたらすのかを回帰モデルなどで推定することには無理がある。なぜなら，説明変数である就労年数に含まれる誤差のために，統計モデルの係数の推定値に偏りがもたらされるからである。

簡単な解決策の１つは，就労年数の末尾０ないし５が階級の中ほどに位置するように就労年数を階級化することである。たとえば，4-7 年，8-12 年，以下，５年間隔で階級を構成する（図２(b)）。階級化によって，もとのデータに含まれる情報が多少なりとも失われる。しかし，それによって，回答誤差の影響を緩和できると同時に，個体間の就労年数の長短の関係がある程度保たれる。また，就労年数０年，１年，２年，３年も別々のグループとして，１つ１つのグループをダミー変数で表せば，就労年数の非線形的な効果も推定できる。就労年数４年から７年までの差が区別できなくなるなどの短所はあっても，推定の偏りはかなり緩和できるだろう。

(3) 集計量を再現する

公的統計ミクロデータの場合，普通は集計用の乗率も提供される。集計用の乗率とは，集計表を作成するときに使われる数値で，通常は，１つ１つの観察値に集計用の乗率を掛けて合計すると，母集団の合計の推定値になる。集計表で公表されている統計においては，集計の過程で外れ値処理が施されることがある。このため，自分で乗率を用いて集計しても，公表されている統計を正確には再現できないことも多い。しかし，おおよそは再現できるはずである。ミクロデータのどの部分がどの変数に対応しているかを確かめることも兼ねて，集計表の一部を再現することを勧める。

4　終わりに

ここで述べた事柄は，ミクロデータ分析に限らず，あらゆる統計分析で最初に行う作業である。少しでも統計分析の経験をお持ちの読者には，至極当然と感ぜられただろう。けれども，上達したからと言って基本を疎かにはできない。特に，ミクロデータは，たとえ同じ名称の調査，たとえば『国民生活基礎調査』であっても，中身が調査年によって異なる。データは生ものと表現され，熟練者ほど基本を丁寧に遂行している。拙稿が，『国民生活基礎調査』を利用したミクロデータ分析の手始めに役立つことを願う。

謝辞　図２は，東京大学社会科学研究所附属社会調査・データアーカイブ研究センター SSJ データアーカイブのリモート集計システムを利用し，同データアーカイブが所蔵する日本版 General Social Surveys〈JGSS-2005〉の個票データを二次分析して得た。また，『国民生活基礎調査』における用語の定義や標本抽出法について，厚生労働省細井俊明世帯統計官からご教示いただいた。記して謝意を表したい。

参考文献

厚生労働省『国民生活基礎調査』
https://www.mhlw.go.jp/toukei/list/20-21.html

さいごう・ひろし　早稲田大学政治経済学術院教授。主な論文に「学歴と寿命」『統計』第 70 巻，第７号，pp. 41-44（日本統計協会，2019 年）など。統計調査論専攻。