

# 調査設計とデータの特性に関する基礎分析

石田賢示

## 1. はじめに

本稿では、2012年度の東北大学教育学部開講科目「統計的調査実習」において実施された調査の設計と、得られたデータの基本特性を整理する。「統計的調査実習」では調査企画、調査票設計、標本抽出、実査、エディティング、コーディング、データ入力、クリーニング、分析、そして報告書刊行に至る一連の過程を、受講者が体得することを目的としている。同時に、得られたデータを用いて社会科学的な命題を検証するために必要な仮説構築の方法、および統計的手法について理解を深めることも目的の1つである。

なお、この実習は2011年度より山形大学地域教育文化学部の社会調査実習と合同で実施されている。報告会なども合同で行うため、受講生にとっては自らの研究が多様な関心を持つ他の学生の目にさらされることになる。このような経験はある程度の緊張感を伴うものであるが、論理的なプレゼンテーションを行えるようになる上でも、多少なりとも役に立っているのではないかと思われる<sup>1</sup>。

以下では調査設計、スケジュールおよびデータの特性について記述する。基本的な構成は昨年度の報告書と同様であるが（三輪 2012）、今回は欠票の分析をシンプルに行った。昨年度の反省にもとづき2012年度は、設計段階での割り当てにより学歴分布の偏りに対処しようと試みたが、成功したとは言い難い結果となった。

## 2. 調査の設計とスケジュール

調査概要は表1に示される通りである。昨年度同様、調査対象者は若年層に限定しているため、対象者の年齢範囲は20歳以上40歳以下に設定している。ただし、調査では職業に関する情報を収集する必要があったため、学生は調査対象から除いている。

調査対象者は、表1の調査会社の協力により、同社のインターネット調査モニターのうち条件に合致する方々に協力を依頼することで選定した。なお、インターネットによる事前調査の段階で、学歴毎に割付を行った。この手続きは昨年度の調査実習での結果を踏まえたものである。昨年度の調査実習データの基礎分析の結果、とりわけ母集団特性からの偏りが大きいのは学歴についてであり、学歴分布の偏りを補正することで職業分布の偏りにもある程度対処できる可能性があるとされていた（三輪 2012: 6）。そこで、2012年度の調査実習では性別、年代と学歴でクロスして母集団レベルでの分布を調べた上で、事前層化を行った。無論、学歴の偏りを補正したとしても得られたサンプルが母集団を代表して

---

<sup>1</sup> 調査実習の過程では、山形大学側の山本英弘准教授には調査企画の段階から多大なご協力とアドバイスを頂いた。また、山形大学側の受講生の方々には山形大学で開かれた最終報告会などでお世話になった。記して感謝の意を表しておきたい。

表 1 東北大学教育学部 2012 年度統計的調査実習 調査概要

|           |   |
|-----------|---|
| 調査名称:     | 若年者のライフスタイルと意識に関する調査  |
| 調査目的:     | 現代日本の若年層に対し、教育、就業、政治に関する意見や行動について質問票調査を行い、それらに関わる社会科学的命題を検証する。  |
| 調査項目:     | (1) 基本属性項目<br>(2) 仕事・収入に関する項目<br>(3) 青少年期の家族構成・家庭環境に関する質問項目群<br>(4) 居住地移動に関する質問項目群<br>(5) 学生時代のアルバイト経験に関する質問項目群<br>(6) 人間関係形成に関する質問項目群<br>(7) 社会意識に関する質問項目群(教育・就業・政治) |
| 調査対象:     | (1) 母集団 学生を除く日本全国の20歳以上40歳以下の男女<br>(2) 標本規模 500<br>(3) 標本抽出法 上記条件を満たすインターネット調査モニターのうち学歴毎に割付を行った上で協力依頼を行い、依頼に応じた先着927名から無作為抽出。   |
| 調査時期:     | 2012年6月   |
| 調査方法:     | 郵送調査法(事前の協力伺いと一部質問はインターネット調査)   |
| 調査実施協力機関: | 楽天リサーチ株式会社  |
| 回収結果:     | 有効回収票数(率) 465(93%)  |

表 2 学歴の分布

|            | 計画サンプル<br>(N=500) | 無作為抽出前<br>(N=927) | H22国勢調査<br>(20歳～39歳) |
|------------|-------------------|-------------------|----------------------|
| 中学校        | 6.8%              | 5.9%              | 5.7%                 |
| 高校         | 40.2%             | 40.5%             | 41.0%                |
| 短大・高専・専修学校 | 23.0%             | 23.8%             | 24.0%                |
| 大学以上       | 30.0%             | 29.8%             | 29.2%                |

※割付の際に用いた国勢調査の学歴分布には、在学者も含まれている。

いるとはまったく言えないわけではあるが、基本属性の極端な偏りを避けることを目的として、今回は事前に層化することを試みた。

上述の手続きを経て事前のインターネット調査を行い、協力依頼およびインターネット調査に応じた先着 927 名のなかから 500 名を無作為抽出した。無作為抽出とはいえ、抽出するためのリストがそもそもインターネット調査モニターからなるものであること、さらにその中で調査依頼に素早く応じた人々に限定されている。本来の母集団は日本の若年層ということになるが、ここでの無作為抽出はその縮図をつくりだすことにはなっていない。この点も昨年度同様であり、留意すべき点である。なお、無作為抽出後の計画サンプル、応募者リスト、そして平成 22 年実施の国勢調査による学歴の分布を表 2 に示した。実査前の時点では、当然ではあるが学歴の偏りはほとんどないといえる。

実査は 2012 年 6 月に、郵送法によって行われた。5 月 23 日に先述の事前インターネッ

表 3 東北大学教育学部 2012 年度統計的調査実習スケジュール

|            |  |
|------------|--|
| 2012年4月10日 | 実習授業開始   |
| 2012年5月14日 | 調査会社へと依頼   |
| 2012年5月23日 | 事前インターネット調査開始                                      |
| 2012年5月29日 | 事前インターネット調査終了、調査協力者リスト納品<br>中間報告会(於:東北大学)          |
| 2012年6月5日  | 質問項目締切・調整  |
| 2012年6月8日  | 調査票最終確認  |
| 2012年6月12日 | 郵送調査開始(封入・投函)<br>回収済み調査票に関してエディティング・アフターコーディング作業開始 |
| 2012年6月25日 | 郵送調査締切   |
| 2012年6月26日 | エディティング・アフターコーディング・データ入力(~7月5日)                    |
| 2012年7月9日  | データクリーニング終了  |
| 2012年8月10日 | 最終報告会(於:山形大学)                                      |
| 2012年8月31日 | 実習レポート締切   |

ト調査が行われ、5月29日に締め切りで、対象者候補リストが納品された。この間調査票設計の作業が並行して進められている。その後、リストから抽出された500名の対象者に対して、6月12日にメール便により調査票が送付された。なお事前にはあるが、対象者には500円分の図書カードを謝礼として調査票、返送用封筒、添書とともに送付している。封入、投函作業は東北大側が300票分、山形大側が200票分分担して行った。

調査対象者からの返送の締め切りは6月25日とされ、データ入力作業の最終日となった7月5日までに届いた調査票を入力した。結果、465名分の調査票を有効票とした(有効回収率は93%)。

この間、返送された調査票のエディティング、アフターコーディング作業も並行して進めた。今年度は居住地移動の質問で市区町村までを尋ねており、それらは日本標準地域コードに即してコード化されている。エディティング、アフターコーディングの作業は授業内外の時間で行われ、授業外では受講生のうちの有志およびTAによって作業が行われた。

一連のスケジュールを表3に示した。<sup>2</sup>表3に記載されたもののほか、この間に仮説、質問項目、分析経過等6度の報告が受講者には課され、社会調査の知識、統計的手法やソフトウェア操作の解説も行われている。

<sup>2</sup> 表3に示される通り、本実習は4月から8月までの半期の間で開講される。参考のために記しておく、1週2コマ分(90分×2)の授業となる。短期間の中に社会調査に関する講義、演習と共に調査設計、実査までをこなす必要がある。そのため、調査設計に関する作業のほとんどの部分は担当教員とTAによってなされる。本実習の中ではランダムサンプリングや調査票設計に関する演習課題が課されるため、それをもって受講学生は調査の過程を「体験」する。しかし実際には、質問項目数、質問内容が対象者集団にとって回答しやすいものかなど、調整すべき点が存在する。時間に余裕をもって調査実習を進められるならば、これらのマネジメントに関わる部分に受講学生がもう少し関わりを持ってよいかもしれないと思われる。

### 3. 回収サンプルの特徴

以下では、前節で述べられた手続きで得られたデータの特徴について基礎分析を行う。インターネット調査モニターを対象とした調査であるため、サンプルに偏りが生じることが避けられない。しかし、どのような偏り方をしているのかを確認することは、今後インターネット調査を利用した調査実習を改善する上での基礎作業として重要である。以下での基礎分析では、2010年10月におこなわれた国勢調査の数値を、基準となるデータとして用いる。調査時点からやや離れているが、全数調査であるということもあるため、国勢調査の結果を用いることとした。

表4は、国勢調査との基本属性分布の比較を行った結果を示したものである。2011年度調査実習では婚姻状況に関する質問がなされていたが、2012年度実習では含まれていない。そのため、性別、年代、地域、学歴、雇用形態、職業に関して比較を行うこととした。

2011年度調査実習データでは「思いのほか属性的な偏りが小さめであった」（三輪 2012: 4）ということであった。しかし、表4をみる限り得られたデータの偏りは、少なくともこれらの属性については性別以外で明確に現れているということが、カイ2乗値による適合度検定の結果からも明らかである。

年代については、20歳から24歳の層の割合が過小となっている。これは、比較基準となっている国勢調査の年代分布には学生も含まれているためであろう。地域については、北海道／東北、中部、関西で過大、関東、九州／沖縄で過小となっている。

より注意すべき偏りは学歴、従業上の地位、そして職業である。まず学歴については、国勢調査の分布に即して事前層化を行っており、実査前の計画サンプルについてはほぼ偏りはないといえる。しかし、郵送調査後の分布では中学校、高校卒の対象者の割合が過小、高等教育以上が過大となっている。事前層化を行ってもなお回収段階で偏りが生じるとなると、これらの層をオーバーサンプリングするという手段などが考えられるが、それによる別種の偏りのリスクが生じる可能性もあるだろう。

雇用形態については、2011年度は顕著な偏りがなかった。しかし、2012年度調査では正規雇用の割合が過小であり、自営業主の割合が過大であった。これは事前のインターネット調査でも同様であった。特に、自営業層での偏りが適合度検定に用いたカイ2乗値に影響している。

職業に関しては、2011年度は職業のアフターコーディングを行っていたので農業を含めた比較が可能であったが、2012年度は農業か否かを識別できる質問ではなかったため、農業は除いている。とりわけ偏りが大きいのは事務、サービス、運輸・保安、そして生産現場・技能である。事務職は過大であるが、残りの職種については過小である。

表 4 平成 22 年国勢調査との基本属性分布の比較

|  |                   | 度数  | 有効%   | H22国勢調査 | カイ2乗 | p値    |
|--|-------------------|-----|-------|---------|------|-------|
| 性別                                       | 男性                | 240 | 51.7% | 49.9%   | 0.6  | 0.432 |
|  | 女性                | 224 | 48.3% | 50.1%   |      |       |
| 年代                                       | 20-24歳            | 33  | 7.1%  | 15.3%   | 35.5 | 0.000 |
|  | 25-29歳            | 138 | 29.7% | 22.0%   |      |       |
|  | 30-34歳            | 108 | 23.3% | 25.9%   |      |       |
|  | 35-40歳            | 185 | 39.9% | 36.7%   |      |       |
| 地域                                       | 北海道/東北            | 58  | 12.5% | 10.5%   | 17.5 | 0.004 |
|  | 関東                | 145 | 31.3% | 36.5%   |      |       |
|  | 中部                | 101 | 21.8% | 17.9%   |      |       |
|  | 関西                | 91  | 19.6% | 16.4%   |      |       |
|  | 中国/四国             | 36  | 7.8%  | 8.1%    |      |       |
|  | 九州/沖縄             | 33  | 7.1%  | 10.6%   |      |       |
| 学歴<br>* 25歳以上                            | 中学校               | 12  | 2.8%  | 5.4%    | 16.4 | 0.001 |
|  | 高校                | 143 | 33.2% | 39.6%   |      |       |
|  | 短大・高専・専修学校        | 121 | 28.1% | 24.4%   |      |       |
|  | 大学以上              | 155 | 36.0% | 30.6%   |      |       |
| 従業上の地位<br>/雇用形態<br><br>* 国勢調査<br>20歳-39歳 | 経営者・役員            | 7   | 1.9%  | 1.7%    | 17.2 | 0.008 |
|  | 常時雇用の正社員・<br>正規職員 | 219 | 60.8% | 66.5%   |      |       |
|  | パート・アルバイト・<br>その他 | 81  | 22.5% | 22.6%   |      |       |
|  | 派遣                | 19  | 5.3%  | 4.0%    |      |       |
|  | 自営業主              | 22  | 6.1%  | 3.1%    |      |       |
|  | 家族従業者             | 10  | 2.8%  | 2.0%    |      |       |
|  | 内職                | 2   | 0.6%  | 0.2%    |      |       |
|  |                   |     |       |         |      |       |
| 職業<br><br>* 国勢調査<br>20歳-39歳              | 専門・技術             | 71  | 20.8% | 18.5%   | 52.0 | 0.000 |
|  | 管理                | 6   | 1.8%  | 0.5%    |      |       |
|  | 事務                | 100 | 29.2% | 21.3%   |      |       |
|  | 販売                | 51  | 14.9% | 15.5%   |      |       |
|  | サービス              | 24  | 7.0%  | 12.2%   |      |       |
|  | 運輸・保安             | 31  | 9.1%  | 5.0%    |      |       |
|  | 生産現場・技能           | 59  | 17.3% | 27.0%   |      |       |

## 4. 欠票の分析

2011年度の調査実習データの基礎分析にもとづき、2012年度は学歴による事前層化を行ったものの、回収段階で中卒、高卒層の未回収可能性が示唆された。回収率自体は93%と高いが、回収状況と基礎属性の関係について一応の確認をしておきたい。

表5は、計画サンプルのなかで調査票を返送しなかった場合を1、した場合を0とする二項ロジット回帰分析の結果である。独立変数には年齢、性別、雇用形態、および学歴を用

表 5 欠票に関する二項ロジットモデルの推定結果

|               | 係数     | 標準誤差 |
|---------------|--------|------|
| 年齢            | -0.03  | 0.04 |
| 女性ダミー         | -0.43  | 0.41 |
| 非正規雇用ダミー      | 0.42   | 0.45 |
| 自営・家族従業・内職ダミー | -1.09  | 1.06 |
| 無職ダミー         | -0.17  | 0.56 |
| 中学校卒ダミー       | 1.34 * | 0.57 |
| 専修学校・高専卒ダミー   | 0.50   | 0.51 |
| 短大卒ダミー        | 0.20   | 0.82 |
| 大学以上卒ダミー      | 0.09   | 0.47 |
| 定数            | -1.74  | 1.18 |
| -2LL(切片のみ)    | 253.64 |      |
| -2LL(モデル)     | 242.81 |      |
| N             | 500    |      |

p < .05

いている。これらの変数の情報は、事前のインターネット調査で尋ねられている質問から得られている<sup>3</sup>。

モデル自体は有意ではないが、推定の結果、中卒ダミーの係数がプラスに有意であった（基準は高卒）。これは、高校卒の調査対象者にくらべて中学校卒の対象者は約 3.8 倍の確率で返送しにくいことを意味する。一方、雇用形態間で返送の有無の確率に差はないといえる（基準は正規雇用）。

## 5. おわりに

以上の基礎的分析の結果は、前年度の反省にもとづき学歴と年代で層別を行なっても、回収段階でやはり偏りが生じてしまうことを意味しているといえよう。また、年代については対象者リストの納品時点で既に偏りがある。事前の割付で年代は 20 代、30 代という区分にした。これは割付を細かく行うほどコストがかかるという理由と、若年中卒者の数がそもそも少なく、各年代前半、後半で分けると条件を満たす対象者が応募しない可能性があるという理由による。その結果年代は大まかな分け方にせざるを得ず、20 代前半の少なからぬ規模は学生段階であり、「20 代」の大半が 25 歳以上の者になってしまうという問題が生じる。

学歴分布の偏りを割り当てによって対処するならば、母集団の分布に即しつつも脱落しやすい層をオーバーサンプリングする手段が考えられるということは、先述の通りである。しかし、調査票を返送しにくいということが何らかの特性を反映した結果であるとするならば、そのような層を過大に抽出することによってデータの他の側面で偏りを生じさせる

<sup>3</sup> 雇用形態は、経営者・役員および正社員を正規雇用、パート・アルバイト・派遣・契約・嘱託を非正規雇用、自営業主、家族従業、内職は 1 つのカテゴリにまとめた。

危険性もありうるだろう<sup>4</sup>。標本設計の工夫のみならず、質問票の構成をよりシンプルにする、質問は明快に、複雑なものを避けるようにするなど、調査票設計の基本により忠実な姿勢が実施側にも求められるのではないだろうか。インターネット調査段階ではなかった学歴の偏りが回収段階で生じるということは、多かれ少なかれ調査票設計上の問題について反省すべきである<sup>5</sup>。

また、年代の偏りについては、非学生という条件を外すことである程度対処可能なのではないかと推察される。ただし、20代前半の層における学生の割合は少なくはなく、有業者を念頭に置いた分析を行う場合には、多くのケースを除外せざるを得ない。この点にも対応するとなると、計画標本規模をさらに大きくする必要が生じる。しかし、それに伴うコスト増大は、金銭的費用のみならず、実査、データ入力段階での受講生、TAの時間と労力の側面など、多岐にわたることが明らかである。対象者を20代後半以降に絞り込むことは対処法の1つであり、最も現実的かつローコストな方法であるかもしれない。

現在のような設計での調査は、この調査実習では3年目になる。年度間で共通の質問を含めることで、データの質を比較しやすくもなるし、データのマージが可能になることで少なくとも教育用データとしてはより使い勝手の良いものになると思われる。調査実習が終わると二度とそのデータが使われなくなる、ということはおそらく多くの社会調査実習の授業で起こっている事態であると想像するが、多くのコストをかけて作り出すデータなのであるから、それらを有効活用できるような方法も今後考えてゆくことが望ましいだろう。

#### 【文献】

三輪哲, 2012, 「調査設計とデータの特性」 東北大学教育学部・教育政策科学研究室編『平成23年度東北大学教育学部・教育学実習 「統計的調査実習」 報告書』, 1-7.

---

<sup>4</sup> そもそも外部妥当性が担保されていないデータではあるが、さらなる偏りを生じさせてしまう可能性がある。

<sup>5</sup> 例年共通して見られるのが、調査票には盛り込むものの最終的なレポートでは使わないような質問の多さである。失敗することも実習で得られる1つの教訓であるという考え方や、受講生にとって質問を考える段階では非常に「思い入れ」のある質問であるように感じられることから、教員、TAの側も質問を厳選することには躊躇しがちであった。その結果、調査票のページ数を増やさずに質問を残そうとすると1ページのなかに質問を詰め込むことになる。質問の複雑さもさることながら、調査票の視覚的な問題も、回答負担に影響をあたえる可能性があるという点は、今後の反省点として挙げておきたい。

