

調査設計とデータ特性に関する基礎分析

濱本真一

(東北大学 大学院教育学研究科)

はじめに

本論では、東北大学教育学部の 2013 年度前期開講科目「教育学実習（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（以下本実習と称す）」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。学生たちは本実習のカリキュラムを通じて、社会調査の実際、および社会調査を用いた論文作成の基本的なスキルを学ぶことになる。

後にも詳しく述べるが、本実習で行われた調査は調査票を用いた標本調査である。標本調査において対象者となる標本が対象とする母集団の適切な縮図でなければならないということは、調査法の基礎としてよく知られている。対象者の選び方は母集団に対する単純無作為抽出が原則とされる。しかし多くの場合、標本調査はその目的や制約によってさまざまな方法で行われ、その方法に応じて得られた標本は何らかの独特な「クセ」を伴うことが多い。本調査においてもそれは例外ではなく、得られたデータの特徴を理解しておくことは、本調査を用いて何らかの社会科学的な知見・命題を導き出す際の留意点を理解することと同じである。本章では、本調査のサンプリングの方法を整理し、さらに得られた回答者の属性の分布、回答率、回答の傾向に関して、それぞれ基礎的な分析を行う。

1. サンプルデザインと調査のスケジュール

調査の概要は表 1 に示したとおりである。対象者は例年通り 20 歳以上 40 歳未満の非学生に設定している。学生は各々の問題関心に沿って質問項目案や分析計画をたてた。付録に記載の調査票は各班の関心を最大限生かして作成されたものである。

対象者の抽出は、まず協力調査会社のモニター登録者のうち、条件に合致する方々に協力を依頼し、1012 名の 1 次協力者サンプルを得た。そこから 500 名を無作為に抽出し計画サンプルとした。

抽出に際して性別・学歴・年齢層による割付を行った。昨年までの調査から、特に学歴や年齢層で大きなサンプルの偏りが生じる可能性が予測されており(三輪 2012, 石田 2013), 基本的な属性周辺分布に極端な偏りを生じさせないための措置である。しかし、これらの割付を行うことは、母集団の代表性を担保することにはならない。想定している母集団(理論母集団)は「全国の若年者」であるが、学生でないという条件のため、多くの 20 代前半の人々は母集団から除外される。年代の分布が国勢調査と異なることは、むしろ当然ともいえる。また、インターネットモニターという調査方法の性格上、実質的な母集団(調査母集団)は、「インターネットモニターに登録している全国の学生でない若年者」である。

表1 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
調査主体	東北大学教育学部・教育学実習／山形大学地域教育文化学部・社会調査演習
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紙調査を
調査対象	理論母集団：調査時点で20～39歳の全国の男女(学生を除く) 調査母集団：上記の条件を満たすインターネット調査モニター
データ数	モニター調査によって抽出した500名に調査票を郵送し、471名からの回答を得た(回収率94%)。
調査時期	2013年6月10日～22日
抽出方法	上記インターネット調査モニターのうち、学歴・性別・年代ごとに割付を行ったうえで協力を依頼。
調査方法	基礎項目についてウェブ上で尋ねたうえで、調査法を郵送
調査項目	1 基本属性項目 2 学歴・職業に関する項目 3 青少年期の家族構成・家庭環境に関する項目 4 青少年期の学習習慣・習い事に関する項目 5 生活習慣・社会参加に関する項目 6 職業に関する認識に関する項目 7 社会・教育に関する意識に関する項目
調査協力機関	楽天リサーチ株式会社

表2 性別×年齢層×学歴のクロス表

年齢性別	国勢調査				計画標本				有効回答			
	中学校	高校	短大・高専	大学・大学院	中学校	高校	短大・高専	大学・大学院	中学校	高校	短大・高専	大学・大学院
20代男性	1.34%	9.09%	2.81%	6.40%	1.40%	8.60%	2.80%	6.20%	1.27%	8.70%	2.97%	6.16%
30代男性	1.96%	12.80%	4.40%	10.92%	2.00%	13.00%	4.40%	10.80%	2.12%	12.95%	4.67%	10.19%
20代女性	1.01%	7.94%	5.95%	5.34%	1.20%	7.60%	6.00%	5.40%	1.27%	8.07%	5.94%	5.31%
30代女性	1.19%	11.64%	10.69%	6.54%	1.20%	11.80%	11.00%	6.60%	1.27%	11.68%	11.04%	6.37%

表3 統計的調査実習のスケジュール

2013年	4月8日	実習授業開始
	5月27日	プリテスト・質問項目調整
	5月28日	調査会社へ依頼
	5月29日	山形大学との合同報告（於：東北大学）
	6月7日	調査票最終盤完成
	6月10日	調査票封入・投かん(郵送調査開始)
	6月22日	調査票回答締切
	6月24日	データ入力開始
	7月1日	アフターコーディング・クリーニング
	7月8日	配布用データ完成
	8月3日	最終報告会（於：山形中央公民館）
	8月23日	実習レポート締切（TAによる点検開始）
	9月30日	最終稿提出

調査母集団が理論母集団から無作為に選ばれていない以上、調査母集団から得られるサンプルには理論母集団からの特性の乖離が生じる。例年同様、留意すべき点である。

割付は性別 2 カテゴリ、学歴 4 カテゴリ、年齢層 2 カテゴリの計 16 カテゴリで分類し、平成 22 年度の国勢調査をもとに行った。国勢調査、計画サンプル、有効サンプルにおける各カテゴリの分布は表 2 に示す通り。計画サンプル、有効サンプルともに国勢調査と比べて大きな偏りは見られない（ χ^2 乗値による適合度検定はいずれも有意でない）。

本実習のスケジュールは表 3 のとおりである。学生は授業内でテーマ設定、仮説構築、

表4 回答者の分布

カテゴリ		実数	比率	国勢調査 H22	χ^2 乗値	p値
性別	1 男性	230	48.9%	48.7%	0.012	0.912
	2 女性	240	51.1%	51.3%		
年齢層	20-24歳	60	12.8%	14.8%	3.444	0.328
	25-29歳	127	27.1%	24.2%		
	30-34歳	124	26.4%	28.0%		
	35-39歳	158	33.7%	33.0%		
婚姻状況	1 未婚	245	52.1%	53.5%	0.343	0.842
	2 既婚	209	44.5%	43.2%		
	3 死別・離別	16	3.4%	3.4%		
従業上の 地位	1 経営者・役員	6	1.7%	1.7%	84.997	0.000
	2 常時雇用	198	56.1%	66.5%		
	3 臨時雇用	73	20.7%	22.6%		
	4 派遣社員・嘱託	39	11.0%	4.0%		
	5 自営業主	27	7.6%	3.1%		
	6 家族従業者	7	2.0%	2.0%		
	7 内職	3	0.8%	0.2%		
職業	1 専門・技術	83	24.3%	18.4%	94.519	0.000
	2 管理	11	3.2%	0.5%		
	3 事務	102	29.8%	20.8%		
	4 販売	43	12.6%	15.5%		
	5 サービス	34	9.9%	12.7%		
	6 運輸・保安	19	5.6%	4.7%		
	7 生産現場・技能	48	14.0%	26.0%		
	13 農業	2	0.6%	1.3%		
学歴	中学校	17	3.6%	5.4%	6.622	0.085
	高等学校	170	36.2%	39.6%		
	短大・専門	129	27.5%	24.4%		
	大学・大学院	153	32.6%	30.7%		

質問項目、分析結果に関するプレゼンテーションを行うのと同時に、封入作業、データ入力などの実査作業も行う。表3にあるようにテーマ設定から郵送開始の期間、またはデータ完成から報告書執筆に至る期間は非常に短い。そのため、授業外の時間に各班が自主的に作業を行うことが求められている。なお、今年度は新しく、授業内に調査票のプリテストを導入した。暫定版の質問紙を試しに回答し、互いに議論しあうことで、質の良い調査項目の作成に寄与したと思われる。500人の調査対象者のうち300名を東北大の学生が、200名を山形大学の学生がそれぞれ封入・投かんし、返送された調査票もそれぞれ入力した。

2. 回収サンプルの特徴

回収されたサンプルの属性の分布を表4に示す。これらの分布が母集団を代表しているものなのかを判断するには、外的な基準が必要である。本論では、やや年度が異なるが、平成22年度の国勢調査を用いる。国勢調査における分布を真の分布と仮定し、本調査で得られたデータの属性との整合性を検証する。割り当てを行った3つの変数のうち、年齢と性別に関しては、およそ国勢調査と変わらない分布が得られた。学歴に関しては適合度検定の結果は10%水準で有意であるものの、おおむね偏りのないデータが得られたとみてよ

いだろう。

従業上の地位は国勢調査と異なる結果となっている。特に国勢調査とのかい離が大きいのは派遣社員・嘱託社員カテゴリである。職業²⁾にも有意なゆがみが生じている。とりわけ管理職および事務職は過剰に、生産現場・労務職は過小にサンプリングしていることがわかる。本実習では例年職業に関する分布のゆがみが大きい傾向にある。インターネット調査の特性としてある程度は仕様のないこととはいえ、職業に関する分析を行う際には、留意が必要である。

3. 欠票の分析

本調査の回答率は 94%であり、例年と同様高い値となっているが、事前のインターネット調査において尋ねた質問群を用いて回答と基本属性の関係について検証しておく。まず、性別・学歴・年齢層を用いた割付と回収率の関係は表 5 のとおりである。最終学歴が中学校の 20 代男性の回収率がやや小さいように見えるが、それほど大きな違いではない。属性ごとにみると、男性の回収率(93.9%)は女性(94.5%)よりやや低く、20 代の回収率(95.4%)は 30 代の回収率(93.4%)よりもやや高い。社会調査の回収率は一般に女性のほうが男性より回収率がよくなり、また年齢が上がるにつれ回収率がよくなる傾向を持つことが知られている(三隅・三輪 2008)。本調査では、性別に関してはそれと同様の傾向が、年齢に関しては逆の傾向がみられたが、どちらも値は高く、その差はごく小さいため、明確な回収率の違いとは言えない。

本調査に対して「返送有=1」「返送なし=0」とする二項ロジスティック回帰を行った結果を表 6 に示した³⁾。2012 年度の調査では、中卒者ダミーの効果が有意であり、「高校卒の調査対象者に比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」という結果であった(石田 2013)が、本調査ではそのような傾向はなく、学歴による返送傾向の違いは認められなかった。一方で雇用形態による返送の傾向においては、無職・その他ダミーにのみ有意な効果が認められた。

回収率が高く、さらに基本的な属性項目による返送傾向の違いは認められないにも関わらず、前節でみたような属性の分布の違いが認められるということは、本調査のサンプルの問題は手続きの部分に帰せられるといえよう。すなわち、理論母集団と調査母集団の乖離、およびモニター登録者のうち先着で選ばれる対象者のサンプリング方法にサンプルの

表5 割り当てごとの回収率

年齢・性別	最終学歴			
	中学校	高等学校	短大/高専	大学/大学院
20代男性	86%	95%	100%	94%
30代男性	100%	94%	100%	89%
20代女性	100%	100%	93%	93%
30代女性	100%	93%	95%	91%

表6 回答ありに関する二項ロジット

	Coef.	Std. Err.	オッズ比
WEB男性	0.396	(0.42)	1.49
WEB年齢	0.003	(0.04)	1.00
WEB非正規	0.147	(0.48)	1.16
WEB自営・内職	1.251	(1.06)	3.49
WEB無職・その他	2.421 *	(1.06)	11.26
WEB中卒	0.159	(1.09)	1.17
WEB短大・専門	0.195	(0.57)	1.22
WEB大卒	-0.515	(0.46)	0.60
定数	2.239 †	(1.25)	9.38

N=500, -2LL=206.59 df=8 chi.sq=14.84*

表7 WEBと質問紙における従業上の地位のクロス表

WEB働き方	質問紙働き方										Total
	経営者・役員	正社員	臨時雇用	派遣社員	契約社員・嘱託	自営業・自由業	家族従業	内職	無職	その他	
経営者・役員	4	1	0	0	0	1	0	0	0	0	6
正社員	2	188	2	0	1	0	0	0	1	0	201
臨時雇用	0	2	61	0	0	0	0	0	1	0	65
派遣社員	0	1	0	19	1	0	0	0	0	0	21
契約社員・嘱託	0	2	3	3	12	0	0	0	1	0	21
自営業・自由業	0	0	1	0	0	25	1	0	0	0	27
家族従業	0	0	1	0	0	0	4	0	4	0	9
内職	0	0	0	0	0	1	0	2	1	0	4
無職：求職あり	0	0	3	2	1	0	1	0	41	0	48
無職：求職なし	0	3	0	0	0	0	0	1	51	1	57
その他	0	1	2	0	0	0	1	0	8	0	12
Total	6	198	73	24	15	27	7	3	108	1	471

不一致率＝ 13.6%

ゆがみを生じる原因があると考えられる。

4. 回答の信頼性に関する考察

以下では、調査票にて得られた回答の信頼性（正確さ）について若干の考察をしていく。前節までに述べたように、本調査では一部の基礎項目をインターネット上で尋ね、そのうえで調査票を郵送している。郵送した調査票の中にも基礎項目は含まれているため、一部の変数については、同一人物に異なる形で同じ質問をしている。性別、年齢、従業上の地位、職業、最終学歴の項目がそれにあたる（インターネットと質問紙で一部選択肢カテゴリの違いがある）。これらの項目は、インターネット上の回答と調査票の回答を比較することができる。このような試みは社会調査法の研究の中でも多くはなく（数少ない例として、林 2004 など）、考察する意義があると考えられる。

表 7～11 は、インターネット質問における質問と、質問紙における質問のクロス表である（学歴は 4 カテゴリに統合）。ウェブ回答と質問紙回答が一致しているときには全度数はクロス表の対角成分にあるはずである。しかし見てわかるように、すべての項目において不一致が生じている。

これらの回答のずれが生じる原因としては、以下のものが考えられる。

- 1) 入力ミス：入力の打ち間違い、わかりづらい回答に対する誤判断等
- 2) 回答者の変化：ウェブ上での回答時から調査票回答時（約 1 か月）の間に状況が変化した等
- 3) 回答者のミス：回答者の誤答、質問意図の誤認等
- 4) 意図的な誤答：見栄や社会的望ましさによる回答者の虚偽回答等
- 5) 代理回答：代理回答者による事実誤認や質問意図の誤認等

このうち、1) の打ち間違いの可能性は低いと考えられる。なぜなら本実習では、入力には細心の注意を払い、1 つの調査票につき 2 名が独立に入力を行い、それらを照合することで入力ミスを発見し訂正しているからである。また、わかりづらい回答（たとえば選択肢の 1 と 2 ともとれる○のつけ方をしているなど）に対しては、入力者が作業監督者に判断を仰ぐような体制になっているが、これらの基礎項目においては、そのような例は報告されていない。

また 2) の回答者の変化は限定的には起こりうることである。特に表 11 に示した年齢は、

表8 WEBと質問紙における職業のクロス表

WEB職業	質問紙職業										Total
	専門・技術	管理	事務	販売	サービス	生産現場・技能	運輸・通信	保安	その他	非該当	
専門・技術	59	1	4	1	0	3	0	0	0	3	74
管理	0	6	1	0	0	0	0	0	0	0	7
事務	4	3	89	1	2	1	0	3	0	5	109
販売	0	0	0	27	0	1	0	0	0	0	30
サービス	4	0	1	8	26	0	0	0	1	1	41
生産現場・技能	4	0	0	2	0	36	0	0	0	0	45
運輸・保安	0	1	0	0	0	1	9	5	1	0	19
その他	12	0	7	4	6	6	2	0	5	96	146
Total	83	11	102	43	34	48	11	8	7	105	471

不一致率= 24.0%

表9 WEBと質問紙における学歴のクロス表

WEB学歴	質問紙学歴					Total
	中学校	高等学校	短大・専門	大学	無回答	
中学校	15	11	2	0	0	28
高等学校	2	158	13	22	0	195
短大・専門	0	1	114	1	0	116
大学	0	0	0	130	2	132
Total	17	170	129	153	2	471

不一致率= 11.5%

表10 WEBと質問紙における性別のクロス表

WEB性別	質問紙性別			Total
	男性	女性	無回答	
女性	239	1	0	240
男性	1	229	1	231
Total	240	230	1	471

不一致率= 0.6%

表11 WEB年齢と質問紙年齢の組み合わせ

一致	460
質問紙=WEB+1	10
質問紙欠損	1
合計	471

不一致率= 2.3%

ウェブ回答時から一つ年齢が上がっているケースが一定数いるのはむしろ自然なことである。また、この期間に転職をしたり、職場で従業上の地位が変化したりすることも考えられる。しかしそれらがどの程度の頻度で起こり、表 7, 8 の不一致率が自然なものなのかの判断は困難である。インターネット上と質問紙上で異なる性別を回答している回答者がわずかではあるが（有効回答の 0.42%）存在する。性別に関しては時間で変わることはないため、この原因はあり得ない。

残りの 3) 4) 5) は回答者に起因するものであり、先の性別回答のずれについては、代理回答によるものなどが考えられる。ただしこれらは特定することまたは調査者がコントロールすることは困難である⁴⁾。

さいごに

本調査の基本的な特性を分析したところおおむね調査計画段階での割り当ては成功したといえる結果となった。しかし、本実習に関して重要な変数である職業に関しては、依然偏りを持った状態であった。本実習における調査は今年度で 4 回目となり、調査計画や調査票設計にも過去の教訓から学んだ改善を施している。回収に関しては著しい返送傾向の違いは見られなくなったものの、いくつかの属性に無視できない分布の偏りが存在していた。本調査の第一義的な目的は調査主体である本実習の受講者に対し、社会調査の実査に触れさせることにある。そのためある程度のサンプル偏りやエラーは、「必要な失敗」

として学生の教育効果を持ちうる。しかしながら、社会調査そのものが現在置かれている状況を省みると、授業における調査といえいたずらに質の悪い調査または調査データを作ることにも慎重にならなければいけない。

さいごに本年度実習を通じての課題をまとめる。まず第1に、標本の設計と抽出である。これは2011年度から共通の課題であり（三輪2012）サンプリングは基本的に教員と調査会社の調整のもとで行われ、学生は「20~39歳の非学生500名」という計画サンプルを所与としたうえで研究を企画する。学生らが標本を設計、または抽出する作業はなく、それゆえ彼らが本調査におけるこれらの「失敗」を意識する機会が乏しい。さらに、調査対象者を主体的に選んでいるわけではないので、「回収率をいかに上げるか」という現在の社会調査における大きな課題を意識する機会にも恵まれない。第2に、コーディングがあげられる。本調査では調査票を見て分かるように、自由記述欄がいくつかあるものの、それらは選択肢カテゴリの中に入らない人の補足的な情報として用いているものがほとんどである。入力からデータの完成までにきわめて短い時間しかないため、膨大な作業量を要する自由記述に対するアフターコーディングはなるべく行わないという方針によるものである。これによって学生たちが調査データを整える作業として最も重要と思われる部分を意識せずに終わってしまうことが危惧される。

以上のような課題を含みつつではあるが、本実習では、学生に対して問の設定からレポート作成、プレゼンテーションに至るまで幅広いスキルを涵養する機会となり、また本調査は全国の若年層に対して青少年期の生活環境、自身・家族の学歴や職業、教育・健康・職業にたいする意識などをとらえた有意義な調査であったといえる。石田（2013）が指摘するように、この手の調査データは授業の終了とともにほとんど「封印」され、二度と使われなくなるというのは非常に惜しまれる。これまでの蓄積も踏まえて、本調査データを有効に活用する方法を考えていくという課題は、本年度も共通である。

【注】

- 1) さらに言えば、モニター登録者のうち、素早く対象者としての条件を理解し、応募するような特性を持った人である。
- 2) 事前の質問には農業は含まれていなかったが、自由記述欄に農業と書かれた回答が複数あったため、新しいカテゴリとして事後に追加した。
- 3) 割り当てインデックスを用いれば、性別×年代×学歴の交互作用を検討したことにひとしくなるが、表5で確認したように、いくつかの割り当てにおいては完全な識別となってしまうため、それぞれの特性を独立であると仮定して投入した。また、働き方の項目においても無職は回収率が100%となっている。無職者をケースから除外するとサンプルサイズが大幅に減ってしまうため、その他を同一カテゴリに含めて推定した。
- 4) とはいえ、質の良い調査データを作成するためには取り組まなければならない問題でもある。特に職業に関する質問などは、回答者が容易にこたえられるように質問文や選択肢に工夫をしなければならない。本調査では国勢調査を始め多くの社会調査が用いている職業分類を用いたが、回答者によってはこの分類自体が回答困難であるという可能性も否定できない。

【文献】

石田賢示, 2013, 「調査データの特性に関する基礎分析」東北大学教育学部・教育政策科学研究

- 室編『平成 24 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.
- 林英夫，2004，『郵送調査法』関西大学出版部.
- 三隅一人・三輪哲，2008，「2005 年 SSM 日本調査の欠票・回収状況の分析」三輪哲・小林大祐
編『2005 年 SSM 日本調査の基礎分析——構造・趨勢・方法』2005 年 SSM 調査シリーズ
1，2005 年 SSM 調査研究会：17-29.
- 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.