分布の偏りと検定力*

太郎丸 博†

分布の偏りと検定力の関係についてい考える。

SSM の 2005 年調査では、本調査とは別に若年を対象に郵送とインターネットの 2 本立てで調査を計画している。

しかし、インターネット調査に対する懐疑は大きいと思われる。この問題を逆の側面から考えてみよう。つまり、郵送のみで若年の非正規雇用や無職について調査する場合、どの程度のケース数が必要かという問題である。オーソドックスな方法では、ケース数が著しくたくさん必要で、オーバーサンプリングすれば、そのような問題を避けられることを示す。オーバーサンプリングの方法の 1 つとして、インターネット調査が選ばれたのである。

1　無作為に抽出した場合

仮に母集団のうち、フリーターが 5%、ニートが 2% いるとしよう。上層の出身者が全体の 25%、残りの 75% が下層の出身者であるとする。 つまり、母集団の全ケースでクロス表を作ったら、セルのパーセント（セル度数÷有効ケース数）は表 1 のようになるとしよう。オッズ比は約 1.9 倍である。上記の数値は、まったくのでたらめというわけではなく、それなりにもっともらしい数値になっている。さて、もちろんわれわれは母集団全体を調査することはできない。このような母集団からランダムサンプリングして、サンプルからクロス表を作る。このときサンプリングの際の偶然で、サンプルから作るクロス表は、上の表のようには必ずしも分布しない。仮に上の表のように

表 1　母集団における分布（架空）

	その他	フリーター	計
上層出身	0.24	0.01	0.25
下層出身	0.69	0.06	0.75
計	0.93	0.07	1

オッズ比 = 1.9

* Wiki に 2006/10/10 に掲載したものを、Wiki サイトの閉鎖に伴い、pdf 版に変更し、内容も微修正した。

† 京都大学文学研究科, tarohmaru.h@hs2.ecs.kyoto-u.ac.jp。

表 2　表のような母集団からサンプルを得た場合の標本サイズと検定力

標本サイズ	500	1000	2000	3000	4000	5000
ランダムサンプリング時の $1-\beta$	0.2	0.4	0.72	0.88	0.95	0.98
オーバーサンプリング時の $1-\beta$	0.75	0.97	1	1	1	1

分布したとしても、ケース数が少なすぎれば、有意な結果とはならない。仮に上の表のような母集団からサンプリングした場合、有意な検定結果が出る確率はどれくらいだろうか。この確率は検定力と呼ばれることがある。当然検定力は標本サイズに依存するので、標本サイズごとにこの確率を計算してみた（有意水準は 5%)。その結果が表 2 の 2 行目である。ちなみに検定力はいわゆる第 II 種の過誤の確率を β とすると、$1-\beta$ であらわせる。なんと 1000 ケースのデータをとっても有意な検定結果が出る確率は 40.4% しかない。4000 ケースをとったところで、ようやく 95% をこえる。つまり、母集団で最初の表のような分布をしている場合、4000 ケースのデータを集めれば、有意な結果がかなり高い確率で出るということである。しかし、問題はそこでは終わらない。ふつう性別や年齢で分けて分析するので、その場合、そのぶんだけたくさんケースが必要になる。つまり、男女に分けるなら、合計 8000 ケース必要ということである。さらに複雑な分析をするならばもっとサンプルが必要になる。無尽蔵に予算があれば別だが、このような調査はあまりにも効率が悪い。

2　オーバーサンプリングした場合

それでは、フリーターとその他が半分ずつになるようにオーバーサンプリングした場合どうだろか。この場合の検定力は表 2 の 3 行目のようになる。1000 ケース集めればじゅうぶんである。この差は大きい。

3　オーバーサンプリングの方法

オーバーサンプリングするためには、対象者の属性を調査への協力を依頼する前に知る必要がある。この例では、あらかじめ対象者がフリーターかどうか分からなければ、オーバーサンプリングできないということである。ここで用いられるのが、エリア・クォータとかランダム・ウォークとかよばれる方法である。調査会社によっていろいろなやり方があるようだが、基本は、指定された世帯に調査員は行き、そこに指定された属性（今扱っている例ではフリーター）の対象者が住んでいるかどうか尋ね、住んでいれば面接または調査票を留め置くというやり方である。ただし世帯は、住宅地図からランダムに選ばれた家屋が指定される場合もあるし、「この地域の世帯ならばどこでも可」といった大雑把な方法もあるので注意が必要。

この方法は、面接調査同様、在宅率の低い層を対象にする場合、ゆがみが大きく出る危険性があ

る。「この地域の世帯ならばどこでも可」といった大雑把な指定をする場合、在宅率によるゆがみはさらに大きくなる危険があるので、さらに注意が必要である。このような問題を避けるためには、対象者の属性を細かく指定することが考えられる。この対策は、調査員の錬度やモラールが高ければ可能であるように思えるが、リスクもある。

4　なぜ郵送＋インターネットなのか

一言で言えば、予算の制約である。2500 万円ぐらい予算があれば、上記のエリア・クォータを選んでいたかもしれない。しかし、予算の制約が厳しいので、エリア・クォータでは、1000 ケースも集まらない。そうするとオーバーサンプリングしたとしても、十分な検定力のあるデータにならない。確かにインターネット調査のサンプルは歪んでいる可能性があるのだが、郵送の結果と慎重に比較しながら使えば大きな失敗はしないだろうし、じゅうぶんな数のデータが格安で得られる魅力は大きい。

また、フリーターのような統計的マイノリティの分析のためには、オーバーサンプリングは有効だが、今回の調査は、必ずしもフリーターの研究ではない。若年層の階層帰属意識が男女でどう異なるかを調べる場合、フリーターのオーバーサンプリングはむしろ邪魔かもしれない。このようなことを考えると、マイノリティに特に注目しない場合は、オーソドックスな郵送調査のデータだけを使ってもらい、マイノリティに注目する場合は、郵送調査＋インターネット調査で分析してもらうのが良いと判断した。