JGSS でみる層化２段抽出法の地点数割当数の精度評価
―JGSS-2000 での割り当てを例にとって―

稲 葉 太 一

（神戸大学発達科学部）

Estimation of Accuracy for Assignment of Number of Survey Points at Two-stage Stratified Sampling on JGSS:
For example, assignment of number of survey points at JGSS-2000

Taichi INABA

By National Social Survey, we often use two-stage stratified sampling that divide whole country into 6 blocks(Hokkaido-Tohoku, Kanto, Chubu, Kinki, Chugoku-Sikoku, Kyusyu) further divide 3 population-size(13-big city, other city, suburban districts) . This paper estimate two effects(block effect and size effect) by assignment of number of survey points on JGSS-2000 put in two-stage stratified sampling. Now, in fact, actual survey is three-stage stratified sampling include minimal survey block. For more accurate estimation, we necessary variation of minimal survey block.

Key words: JGSS, two-stage stratified sampling, estimation of accuracy

全国調査において、全国を 6 つのブロック（北海道・東北、関東、中部、近畿、中国・四国、九州）に分け、その各々について更に 3 つの人口規模（13 大都市、その他の市、郡部）に分類した層化 2 段抽出法は、比較的多く用いられている。この報告では、層化 2 段抽出法で実施されている JGSS-2000 での調査地点数の割り当て数を例に、そのブロックと人口規模の変動の評価を行う。更に、実際の調査は調査区による変動を考慮に入れた 3 段抽出である。精度評価には、この調査区による変動を調べる必要があることが示唆された。

キーワード：JGSS、層化２段抽出法、精度評価

１．はじめに

全国社会調査において、層別２段サンプリングが一般的に用いられているが、その精度評価は、大変困難である。しかし、実際の年齢分布との乖離がある場合、調査自体の信頼性を損なう可能性があることから、多くの調査では、調査時点の国勢調査による年齢分布（推定値）と比較検討を行なっている。また、調査地点数や各地点での調査人数は、その調査に掛かる費用と調査結果の精度との兼ね合いで決められるべきものであるから、より高い精度での評価の必要性はある。実際に各ブロックでの調査地点数を割り当てる際に、割り当ての約１年後に実際の調査が実施されるため、１年後の人口分布（主に年齢分布）を予測することが必要である。稲葉[1]は、このような状況に対して、人口変化率（死亡率と国外移動率の併せた概念）という考えを用いて一つの方向性を提案している。また、稲葉[2]～[5] では､単純ランダムサンプリングを仮定して､年齢分布のズレを評価している。これは管理図の手法を用いてデータのばらつきを評価しているが、実際に行なわれているのは層別２段サンプリングであることや、そのサンプリング単位における均一性や独立性を仮定していることで、現実とのズレがまだ無視できない可能性が懸念される。

この報告では、渡邉[6]での議論を用い、まず層別２段サンプリングの構造を前提とし、最小サンプリング単位内（調査地点内）の均一性と独立性は仮定して精度評価を試みる。ただ､渡邉等[7]で述べられているように､実際の社会調査の状況では種々の値が不明確であることから、割り当ての是非を検討できるような精度の高い議論は望めず、概略を述べるに留まる。

２．従来からの検討内容

稲葉[2]では､単純ランダムサンプリングであると考えてポアソン分布を仮定し､荒い近似としてデータの年齢分布のズレを評価している。この結果､80 代の割り当て数における若干のズレが示唆された。ここで評価の基礎となっているのは、年代の多重性を考慮して有意水準を小さくした「3σルール」である。20 代から 80 代まで、7 つの年代があり、これらのすべてを別々に検定すると、各々を有意水準 5%で検定しても、全体としての有意水準は 5%よりもずっと大きくなってしまう。そこで、一つ一つの検定における有意水準を低めに設定し、全体としての有意水準が 5%を超えないように考えられたのが「3σのルール」である｡これは､標準正規分布に従う確率変数の絶対値が 3 を越える両側確率 0.0027 の有意水準で、個々の検定を行っていることに相当するため、7 つの年代で検定を繰り返しても、ボンフェローニの不等式より全体の有意水準は 0.0027×7=0.0189 と 5%以下が保証されている。

また更に、上記の検討の２つ目の仮定として、データがポアソン分布に従うという仮定が設定されている。より丁寧に言えば、この仮定は明らかに不正確である。というのは、データは層別した後に、２段サンプリングの手法で得られており、調査区での 20 代の比率

が変動していることも併せて考えると、均一な集団からのランダムな標本であるとは考えがたい。しかし、他に有力な方法も容易には存在しないので、現実的に、割付の概略を知る目的で参考的に計算されている値と捉えることが適切であろう。

３．本来のデータの構造式

この報告では、前節で議論した従来からの評価を更に発展させ、実際に実施されている層別２段サンプリングの構造式を立て、単純ランダマイズを仮定する場合より精度の高い評価を試みる。この解析では、選ばれた調査区の比率の偏りは捉えきれないものの、最初の２段に層別してサンプリングすることの影響は評価できる。

以下ではJGSS-2000 を例に、全国を６ブロック（北海道・東北、関東、中部、近畿、中・四国、九州）に分け、更に市郡規模で３つ（13 大都市、その他の市、郡部）に分類し、調査地点数を割り付ける状況を考える。地域（ブロック）間の違いを要因 A，市郡規模（サイズ）間の違いを要因 B，各調査地点の影響を要因 C とおく。第 i 番目の地域で、第 j 番目の市郡規模で、第 k 番目の調査地点での 20 代の真の比率を P_{ijk} で表すことにする。これらの要因を考慮した解析方法をとして、本来のデータの構造式は、最も単純に考えれば、以下のように設定することができる。

$$P_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\ \beta)_{ij} + \gamma_{ijk}, \quad i = 1,,,6;\ j = 1,2,3;\ k = 1,,,,n_{ij}$$

$$\sum_{i=1}^{6} n_i.\alpha_i = 0, \quad \sum_{j=1}^{3} n._j \beta_j = 0, \quad \sum_{i=1}^{6} n_i.(\alpha\beta)_{ij} = \sum_{j=1}^{3} n._j (\alpha\beta)_{ij} = 0, \quad \gamma_{ijk} \approx N(0, \sigma_{ijk}^2)$$

ここで、α_i は第 i 番目の地域の影響（要因 A）、β_j は第 j 番目の市郡規模の影響（要因 B）、$(\alpha\beta)_{ij}$ は、これらの交互作用 $A \times B$ とする。また γ_{ijk} は誤差項であり、ある調査地点の影響（要因 C）も含んでいる。通常の全国的社会調査では、調査地点が非常に多数であり再現性のある結論を必要としていないので、γ_{ijk} はランダムであると考えるのが適切である。

なお、上記の構造式では誤差項に正規性を仮定しているが、今回のようなバラツキを荒く捉えたいといった目的で最初に分散の検討を行う段階では、正規性の仮定は必ずしも必要ではない。そこで、この誤差項の分散について、P_{ijk} を真の比率とする２項分布によって計算し、$\sigma_{ijk}^2 = P_{ijk}(1 - P_{ijk})$ のように評価することも可能である。

４．理論と実際の違い

この報告では、原則として正規対象の人数についてのみ検討を行なう。これは、予備対象者が、正規対象の欠票が発生した地点に限られるため、欠票の起こる理由として考えら

れている転居や住所不明が、比較的都市部に多いという背景がある。つまり、予備対象を含めてデータに偏りを生じさせるよりも、正規対象のみに限れば純粋に割り付けのズレを評価できると考えられる。ただし、JGSS-2000 では、割り当て年齢ミスによって正規対象が 14 名になっている地点が 2 つある。これらの地点に関しては、予備対象の 1 番目の年齢をデータとして含めて解析する。

また、ある調査地点での 20 代人数の従う分布に、何らかの理論的分布を当てはめようと考えると、2 項分布以外を仮定するのは難しい。しかし、2 項分布が必ずしも正しいと言える訳ではない。実際に例えば、ある調査地点が会社の独身寮に当ったとすると、比較的 20 代の多いグループとなる。これは、調査地点の中にも色んな集団があり、地点内の均一性が保証されていないことを意味する。ただ、これをモデル化することは、現状では困難であるし不可能であると思われる。調査地点内の均一性についての情報をどのように捉えれば良いのかは今後の課題であるといえよう。

この報告では、今後地点間の変動という言葉で、これらの変動も含めて荒く捉える事にする。現状では、ある調査地点内での均一性は仮定せざるを得ない。当然のことながら、調査地点内の比率の変動がある場合は、地点内のバラツキが今回の計算結果より大きくなることに注意してほしい。

５．地域間の違いと市郡サイズによる違い

３節では、本来のデータの構造式を示した。しかし、比率のような範囲を持った値を目的変数とした回帰分析や分散分析では、通常目的変数をそのまま分析することはない。構造を考えていく上で、要因の加法性が不可欠であるからである。この節では、比率の要因分析で定番となっている「ロジット解析」を行う。以下に、ロジット変換の計算式と、この報告で用いたデータの構造式を示し、分析結果を示す。

まず、分析に用いたデータを明確にする。３節で述べたように、地域間の違いを要因 A(block), 市郡サイズの違いを要因 B(size) とおく。地域が i 番目、市郡サイズが j 番目、第 k 番目の調査区での、調査対象者数 15 人の内、$x_{ijk}\ (k=1,,,n_{ij})$ 人が 20 代であるというデータを考える。ここで n_{ij} は、第 i 地域の第 j 市郡サイズにおける調査区の数である。これらを比率のデータと考えて、ロジット変換を施す。ロジット変換は、次の式で計算される。

$$L_{ijk}=\ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right)$$

ここで、$p_{ijk}=\dfrac{x_{ijk}+0.5}{15+1}$ とする。20 代の比率 p_{ijk} のロジット変換値 L_{ijk} に関する、次の二元配置分散分析モデルを考える。

$$L_{\mathrm{ijk}} = \ln\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \mu + a_i + b_j + (ab)_{\mathrm{ij}} + \varepsilon_{ijk}\ , \quad \mathrm{i}=1,,,6;\ \mathrm{j}=1,2,3;\ \mathrm{k}=1,,,,\mathrm{n}_{\mathrm{ij}}$$

なお、変換後のデータに関して、ロジット変換後の分散を、p_{ijk} の変動幅を考慮し、$w_{\mathrm{ijk}} = (15+1)p_{ijk}(1-p_{ijk})$ を重みとする重み付け最小二乗法を考える。もちろん、単純に L_{ijk} を生データだとして、重み無しの分散分析を行うことも考えられるところである。というのは、20 代の比率が大きく変動することは考え難いことと、地点内の均一性も保証されていない状況で大掴みな結果を見たいからである。しかし、一方では、地区や市郡サイズごとの調査地点数 n_{ij} も影響があると考えられるため、この報告では、重み付き平方和を採用し、以下の式で計算する。

$$\mathrm{S}_{\mathrm{A}} = \sum_{i=1}^{6} w_{i}..(\bar{L}_{i}.. - \bar{L})^2\ , \qquad \mathrm{S}_{\mathrm{B}} = \sum_{j=1}^{3} w_{.j}.(\bar{L}_{.j}. - \bar{L})^2\ ,$$

$$\mathrm{S}_{\mathrm{AB}} = \sum_{i=1}^{6}\sum_{j=1}^{3} w_{ij}.(\bar{L}_{ij}. - \bar{L})^2\ , \quad \mathrm{S}_{A\times B} = \mathrm{S}_{AB} - \mathrm{S}_{A} - \mathrm{S}_{B}\ .$$

ただし、添え字のドット（.）は、その添え字に関する和を表わし、

$$\bar{L} = \sum_{i=1}^{6}\sum_{j=1}^{3} w_{ij}.L_{\mathrm{ij}}./\sum_{i=1}^{6}\sum_{j=1}^{3} w_{ij}.,, \quad \bar{L}_{i}.. = \sum_{j=1}^{3} w_{ij}.L_{ij}./\sum_{j=1}^{3} w_{ij}.$$

などとする。

表１　各要因の分散分析表

要因	平方和	自由度	$\chi^2(0.05)$	p-値
A(block)	5.288	5	11.07	0.382
B(size)	8.824*	2	5.99	0.012
A×B	9.133	10	18.31	0.520

結論としては、JGSS-2000 の場合、20 代の比率に関しては、市郡規模は５％有意であったが、地域間、地域と市郡規模の交互作用は５％有意ではなかった。

６．まとめと今後の課題

20 代の比率の分析に限った話であるが、市郡規模のみが影響があった。渡邉[6],[7]によれば、調査区内での変動は完全には捉えきれていないものの、調査区内の 20 代の比率の変動を別のデータから推測することで「20 代の比率に関する推定値の分散」への影響を検

討することができる。その結果によれば「推定値の分散」への影響は、概ね数%から10%程度であった。このことが、割り当てに何らかの影響がある可能性は捨てきれない。また、20代のデータでの分析ではなく、80代のデータでそれらの影響を見ると結論は変わる可能性もある。今後は、渡邉[7]にも記述のある、調査区数や各調査人数などの全データを入手することを目標にすることを考えていきたい。

[Acknowledgment]

日本版General Social Surveys（JGSS）は、大阪商業大学比較地域研究所が、文部科学省から学術フロンティア推進拠点としての指定を受けて（1999-2003 年度)、東京大学社会科学研究所と共同で実施している研究プロジェクトである（研究代表：谷岡一郎・仁田道夫、代表幹事：佐藤博樹・岩井紀子、事務局長：大澤美苗)。東京大学社会科学研究所附属日本社会研究情報センターSSJ データアーカイブがデータの作成と配布を行っている。

[参考文献]

[1] 稲葉太一，2002,「全国調査での層化２段抽出法における最適な地点数割り当て法－JGSS-2000 の地点数割り当ての評価」『日本版 General Social Surveys 研究論文集 JGSS-2000 で見た日本人の意識と行動』p.185-192.

[2] 稲葉太一, 2002,「回収率ならびに欠票の分析」『日本版 General Social Surveys JGSS-2000 基礎集計表・コードブック』pp.11-14.

[3] 稲葉太一, 2003 「回収率ならびに欠票の分析」『日本版 General Social Surveys JGSS-2001 基礎集計表・コードブック』pp.11-14.

[4] 稲葉太一, 2004,「回収率ならびに欠票の分析」『日本版 General Social Surveys JGSS-2002 基礎集計表・コードブック』pp.15-18 .

[5] 稲葉太一・保田時男, 2005,「回収率ならびに欠票の分析」『日本版 General Social Surveys JGSS-2003 基礎集計表・コードブック』.

[6] 渡邉真哉，2002,「全国社会調査の精度評価」神戸大学総合人間科学研究科、修士論文.

[7] 渡邉真哉・稲葉太一, 2005,「全国社会調査における２段サンプリングの精度評価法の提案」発達科学研究紀要、投稿準備中.