2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

社会調査における回答者属性データの地域集計

01102840　会津大学　出水田智子　IZUMITA Tomoko

1. はじめに

世論調査などで社会的な集団の特性、意識、行動などを標本調査として実施するときの中心的な課題は、調査対象者の回答比率を通してその背後の母集団全体の比率を推定することである。このとき必要なことは、標本から得られた結果から母集団の統計値が精度よく推定されることである。

このような推定比率の精度を決定するのは、調査対象者の抽出方法によって生じる母比率と標本比率の差だけではない。調査の精度は母比率と回答比率の差で評価されるため、むしろ回収率の低さや不完全な回答結果あるいは作業ミスなどの影響による標本比率と回答比率の差の開きが影響する。さらに結果を比率で表現するために、1集計単位あたりの標本数の大きさやばらつきによる影響も無視できない時がある。

そこで本稿では、分析結果に含まれる調査誤差のうち、回答者属性データの地域集計作業が母比率推定に与える影響について計算例を基に検討する。

2．調査誤差の定義

調査対象全体の母比率推定における調査誤差は、以下のように標本誤差 ε_S と非標本誤差 ε_R の和で定義される。

$$\begin{aligned}\varepsilon_0 &= p - \dot{p} \\ &= (p - p') + (p' - \dot{p}) \\ &= \varepsilon_S + \varepsilon_R \end{aligned} \quad (1)$$

ただし p は母比率、p' は標本比率、$\dot{p}$ は回答比率である。母集団数、標本数、回答数をそれぞれ N、n、t、その中の各個体の値を θ_i (i=1, …, N)、θ'_j (j=1, …, n)、$\dot{\theta}_k$ (k=1, …, t)、とすると、

$$p = \Sigma\theta_i / N$$
$$p' = \Sigma\theta'_j / n$$
$$\dot{p} = \Sigma\dot{\theta}_k / t \quad (2)$$

となり、各集団におけるその値の平均に等しくなる。

集計誤差は全体を一つとして求めた(1)の結果と、異なる単位で集計した結果を比較することで求めることが出来る。そこで、母集団、標本、回答を共通な属性によってL個の層に分割し集計することをここでは考える。h =1, …, L のとき、層別の母集団数、標本数、回答数を N_h、n_h、t_h、その中の各個体の値を θ_{hi}、θ'_{hj}、$\dot{\theta}_{hk}$ とすると、集計単位別の各集団の比率はそれぞれ

$$p_h = \Sigma\theta_{hi} / N_h$$
$$p'_h = \Sigma\theta'_{hi} / n_h$$
$$\dot{p}_h = \Sigma\dot{\theta}_{hi} / t_h \quad (3)$$

となり、先ほどと同様に各層の値の平均に等しくなる。ここで母集団における各層の構成比率を $w_h=N_h/N$、標本集団における構成比率を $w'_h=n_h/n$、回答集団における各層の構成比率を $\dot{w}_h=t_h/t$ とおくと、集計後に算出される母集団、標本、回答における集計による比率の差は

$$\varepsilon_a = p - \Sigma w_h p_h$$
$$\varepsilon'_a = p' - \Sigma w'_h p'_h$$
$$\dot{\varepsilon}_a = \dot{p} - \Sigma\dot{w}_h\dot{p}_h \quad (4)$$

となる。

そこで(1)と(4)から、集計化による誤差項 ε_a が考慮された調査誤差の推定モデル ε を以下のように定義する。

$$\begin{aligned}\varepsilon &= |(p - \Sigma w_h p_h) - (p' - \Sigma w'_h p'_h)| \\ &\quad + |(p' - \Sigma w'_h p'_h) - (\dot{p} - \Sigma\dot{w}_h\dot{p}_h)| \\ &= (p - p') + (p' - \dot{p}) + (\Sigma\dot{w}_h\dot{p}_h - \Sigma w_h p_h) \\ &= \varepsilon_S + \varepsilon_R + (\Sigma\dot{w}_h\dot{p}_h - \Sigma w_h p_h) \\ &= \varepsilon_S + \varepsilon_R + \varepsilon_a \end{aligned} \quad (5)$$

3．計算例

使用する調査データの概要は以下の通りである。

- 調査目的：会津若松市市民の生活意識調査
- 標本抽出方法：系統抽出法(7つの初期値)
- 抽出枠：住民基本台帳記載の20歳以上の男女
- 母集団数：91420(=N)
- 標本数：2037(=n*)
- 調査方法：無記名回答による郵送
- 調査時期：2001年12月

回収された解答用紙(913)のうち、回答者属性データが網羅されている有効回答(n=787)を対象に、性別や年齢など回答者属性データに関する比率を異なる2つの地域スケール（郵便番号別、地区別）で集計したときの推定値と真値の差を求めた。詳細な計算結果は当日紹介する。

今回利用したデータは地域別世帯順に並んだ抽出枠から等間隔に抽出したため、母数が不明な場合は理論的な式を使った誤差推定ができない。今回は行政の統計で確認した正確な母比率を用いて誤差を得た。念のため事後的に、ランダム・サンプリングの式に標本比率をあてはめて、近似的な区間推定も行った(図1)。

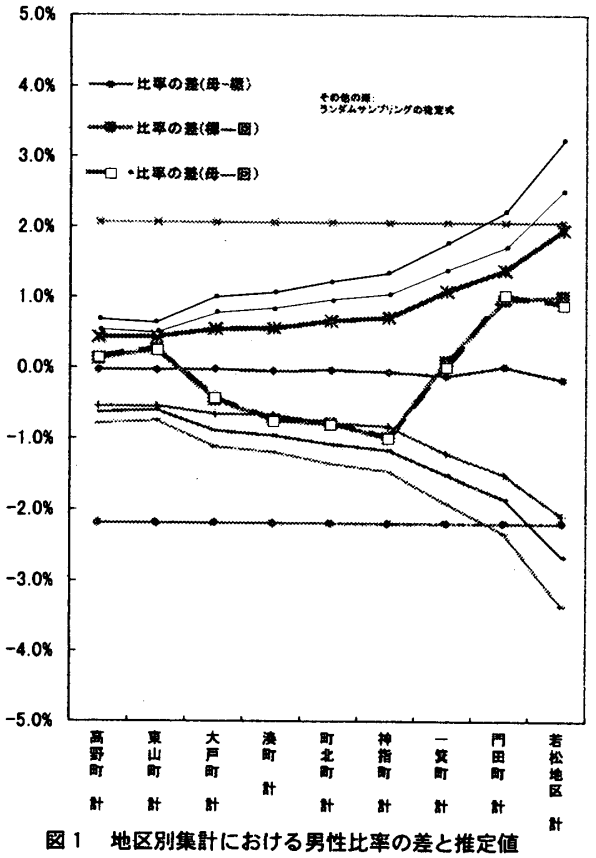

図1　地区別集計における男性比率の差と推定値

集計単位別の母比率と、回答比率がわかれば(5)式によって無回答などによる誤差と集計誤差を区別できることが確認された。

4．おわりに

本稿では、標本誤差ε_Sがある程度わかっているときに非標本誤差ε_Rと集計誤差を区別して求めるための評価式を導出した。今回データが限られたため、十分な検討ができなかったので、今後さらにいろいろなデータで検証作業をおこなう予定である。分散に基づく標本誤差推定ではε_Sの符号が不明なためにそれぞれの絶対値の和をとる場合が多い。このため調査誤差は(1)よりも大きな値で推定されているが、むしろ予測不可能なε_Rの評価が問題である。本来の調査目的である母集団の真の値（例えば母比率）が既知でなければ厳密な意味でε_S、ε_R、ε_aの3種類の誤差を回答結果から分離することはできないが、今回の計算例のように補助情報などによって真の値や標本誤差が既知の回答者属性データで非標本誤差と集計誤差を分離することは可能である。少なくとも回答者属性に関するクロス分析結果を地域全体に一般化する場合や、リサンプリング標本精度の評価において利用可能であろう。今後は居住地情報の集約によって生じる集計誤差が調査結果にどのように影響するか、空間データ的な視点からを検討したい。

参考文献

[1] NHK放送文化研究所世論調査部編(1996)：「世論調査事典」，大空社.

[2] 会津大学文化研究センター(2002)：会津若市における生活と文化に関するアンケート調査—基礎集計—．会津大学文化研究センター研究年報第8号，pp. 92-158.