多重寸法指標のノンパラメトリック推定

岡山商科大学 経済学部　佐井 至道

1　はじめに

官庁統計調査などの標本調査で得られる個票データを公開する際のリスク評価法として，個票データから得られる標本寸法指標を基にした母集団寸法指標の推定について，これまでも議論が重ねられてきた．

個票データは，個体（個人，事業所など）ごとに調査項目の結果の値をレコードとして並べたものである．調査項目のうち，年齢や性別のように，個体を特定するために用いられる項目をキー変数と呼ぶが，すべてのキー変数の組み合わせによってセルが構成されていると考える．例えば年齢と性別のみがキー変数の場合,「40 歳男性」などというセルが構成される．そのセルに含まれる個体数が l のとき，そのセルをサイズ l と呼び，そのようなセルの数を母集団では S_l と表して母集団寸法指標と呼ぶ．同様に標本のサイズ l' のセル数を $s_{l'}$ と表して標本寸法指標と呼ぶ．個票データが標本調査で得られている場合，標本寸法指標が観測され，その情報に基づいて母集団寸法指標を推定するが，特に S_1 など小さいサイズの頻度の推定が重要である．

2　多重寸法指標の導入

個票データは 1 回の調査分についてまとめられているのが一般的である．パネル調査などの継続調査において同一個体が数回の調査でサンプリングされている場合でも，数回分のレコードをリンクした個票データを作成すると，リスク評価の方法は通常の個票データの場合と同様となる．

ここで新たに次のようなリスク評価方法を提案する．

同じ項目に関する数回にわたる継続調査があり，個体はその都度，独立に非復元単純無作為抽出されていると考える．この設定は多くの官庁統計調査に当てはまる．例えば1ヶ月に 1 回ずつ m ヶ月にわたって同じ標本調査が独立に行われたとする．i ヶ月目における母集団の大きさを N_i，標本の大きさを n_i とし，抽出率を $\lambda_i = n_i/N_i$ とする．例えば年齢と性別のみがキー変数の場合，各月の個票データから「40 歳男性」のセルが 1 つずつ構成されるが，このように母集団において各月のサイズが $\boldsymbol{t} = (t_1,\ t_2,\ \ldots,\ t_m)$ であるセル数を $S_{\boldsymbol{t}}$，個票データ（標本）において各月のサイズが $\boldsymbol{t}' = (t'_1,\ t'_2,\ \ldots,\ t'_m)$ であるセル数を $s_{\boldsymbol{t}'}$ とし，すべてのサイズの組み合わせに対する組を，それぞれ $\boldsymbol{S}$ と $\boldsymbol{s}$ で表す．なお，$t_i, t'_i = 0, 1, \ldots, L_i$ とする．ただし L_i は i ヶ月目におけるセル内の個体数の最大値である．

非復元単純無作為抽出によって標本多重寸法指標 $\boldsymbol{s}$ が得られたときの母集団多重寸法指標の尤度関数は,

$$\mathrm{L}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) = \frac{1}{\prod_{i=1}^{m} {}_{N_i}C_{n_i}} \sum_{\mathcal{C}_2} \prod_{\boldsymbol{t}} \Big\{ \frac{S_{\boldsymbol{t}}!}{\prod_{\boldsymbol{t}'} k_{\boldsymbol{t},\boldsymbol{t}'}!} \cdot \prod_{\boldsymbol{t}'} \big(\prod_{i=1}^{m} {}_{t_i}C_{t'_i} \big)^{k_{\boldsymbol{t},\boldsymbol{t}'}} \Big\} \tag{1}$$

と書くことができる．ただし，$k_{\boldsymbol{t},\boldsymbol{t}'}$ は，母集団においてサイズが $\boldsymbol{t}$ であるセルのうち，標本においてサイズが $\boldsymbol{t}'$ と観測される数であり，$\mathcal{C}_2$ は $\boldsymbol{S}$ から $\boldsymbol{s}$ が生成されるような $k_{\boldsymbol{t},\boldsymbol{t}'}$ についてのすべての組み合わせを表す．

通常の母集団寸法指標のノンパラメトリック推定と同様に，尤度関数は

$$\mathrm{L}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) = \frac{1}{\prod_{i=1}^{m} {}_{N_i}C_{n_i} \lambda_i^{n_i}(1-\lambda_i)^{N_i-n_i}} \cdot \prod_{\boldsymbol{t}'} \frac{e^{-\mu_{\boldsymbol{t}'}}\ {\mu_{\boldsymbol{t}'}}^{s_{\boldsymbol{t}'}}}{s_{\boldsymbol{t}'}!} \tag{2}$$

とポアソン近似できる．ただし

$$\mu_{\boldsymbol{t}'} = \sum_{\boldsymbol{t}(\geq \boldsymbol{t}')} \left\{ S_{\boldsymbol{t}} \cdot \prod_{i=1}^{m} {}_{t_i}C_{t'_i} \lambda_i^{t'_i}(1-\lambda_i)^{t_i-t'_i} \right\} \tag{3}$$

である．

3　実データへの適用と問題点

アメリカ合衆国における 1990 年と 2000 年のセンサスの 1%抽出個票データを用い，提案した方法について検討する．ここでは，ワシントン州の 20 歳以上の個票データを母集団と考え（$N_1 = 34542$, $N_2 = 41957$），各年のデータから抽出率 1/2 でサブサンプリングしたデータを標本と考える（$n_1 = 17271$，$n_2 = 20979$）．また，キー変数として 2 回の調査で共通している 12 項目（年齢，実子，血縁の子，性別，結婚，通勤手段，通勤時間，職業，就業，労働週数，週労働時間，年間収入）を考える．なお，このデータには既にスワッピングなどの秘匿処理が施されているため，ここでの目的はデータのリスク評価そのものではなく，推定の良さを見ることである．

図 1 に母集団と標本におけるセルのイメージを示す．縦に並ぶセルは同じキー変数の組み合わせであるが，入っている個体は一般的に年ごとに異なる．

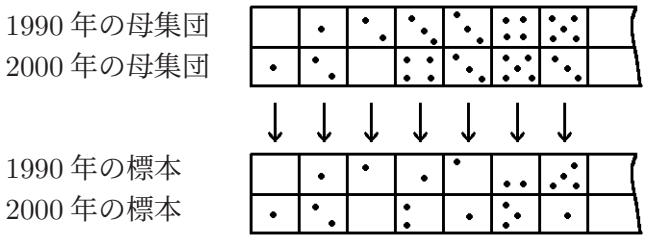

図 1: 継続調査におけるセルのイメージ

推定を安定させるために，母集団寸法指標に対して制約条件をおかなければならない．この検討では，$\boldsymbol{S}$ のすべての要素が非負など，通常の母集団寸法指標のノンパラメトリック推定に準じた 4 種類の制約条件をペナルティー関数に置き換えて対数尤度関数に取り込み，勾配法により探索的に条件付き最大尤度を求めたが，推定結果は通常の母集団寸法指標の推定と比較して良くなかった．

この例では調査間隔が 10 年と長いため，例えば $S_{1,45} = 1$ のように各年でサイズが大きく異なるセルが多数存在することが原因の一つと考えられ，制約条件の追加を含めて更なる検討が必要である．