標本調査法入門

汪　金芳
千葉大学　大学院自然科学研究科

平成 17 年 5 月 16 日

目 次

1　標本調査とは何か　2

2　確率抽出に基づく統計的推測　2

3　単純無作為抽出　4

4　層化抽出　5

5　比推定量　7

5.1　単純無作為抽出 8

5.2　層化無作為抽出 9

6　標本誤差のジャックナイフ推定　10

1　標本調査とは何か

ある集団についての特徴を知るのに，コストや時間的制約などから，目的集団に属する一部の対象のみを調べ，関心のある特徴を推計する統計学の方法が**標本調査法** (sample survey method) である．

標本調査の例として，官庁統計調査（たとえば労働力調査），世論調査（たとえな内閣支持率調査），市場調査（たとえばブランド志向調査），等があげられる．1981 年に行われたフランス大統領選挙におけるル・モンド紙による世論調査が有名である．同年 4 月 27 日〜28 日に世論調査が行われ，社会党の押すミッテラン候補支持と独立共和党の擁護するジスカールデスタン候補支持が，それぞれ 51.5% と 48.5% であった．同年 5 月 10 日の全有権者の投票結果，ミッテラン支持が 51.75%，ジスカールデスタン支持が 48.24%，という驚くべきほど一致する結果がでた．

良すぎた例かもしれないが，この例からも分かるように，多くの場合，集団全体の特徴を掴むため，国勢調査のようにすべての対象を調査する**全数調査** (complete survey，**悉皆** (しっかい) **調査**ともいう) を行う必要はない．

標本調査における目的集団を**母集団** (population) といい，抽出された対象の全体を**標本** (sample) という．また推計すべき母集団の特徴を**母数** (parameter) とよび，標本から構成される母数に対する推計値を**推定量** (estimator) とよぶ．言うまでもなく全数調査とは異なり，標本調査における推定量に誤差が伴う．

標本調査法の目的は，現実的で適切な標本抽出をするための標本設計を行い，それに応じた精度の高い推定量を構築することにある．

適切でない標本設計に基づく調査は時として危険である．1936 年のアメリカ大統領選挙には，共和党からランドン氏が，民主党からはルーズベルト氏が立候補した．リテラシィ・ダイジェスト誌は，200 万人規模の調査を行い，ランドンの勝利を予言した．結果はルーズベルトが当選したのである．廃刊に追い込まれた同誌の失敗の原因は，調査の対象が同誌の購読者と電話の保有者に限定したことにある．当時，このような人たちは高所得者層で共和党支持者である傾向があったからである．最近インターネット調査が多用されるようになり，特にこの種の注意が必要であろう．

2　確率抽出に基づく統計的推測

N 個の要素からなる母集団 U から, 大きさ n の標本 s を抽出することを考える．$f = n/N$ を**抽出率** (sampling fraction) という．通常，N は既知で，また n を調査前に決めておく場合が多い．すべての標本 s の集合 S を**台** (support) と呼ぶ．標本 s が決められた確率によって抽出される方法を**確率抽出** (probability sampling) といい，確率分布

$$p(s) = P(s \text{ を抽出}), \quad s \in S \tag{1}$$

を**標本設計** (sampling design) という．任意の $s \in S$ に対して，標本 s に含まれるすべての要素が異なる抽出法を**非復元抽出** (sampling without replacement) といい，ある $s \in S$ に対して，同じ要素が 2 回以上出現する抽出法を**復元抽出** (sampling with replacement) という．

標本設計 (1) のもとで，母数 θ に対する推定量 $\hat{\theta}$ が

$$E(\hat{\theta}) = \theta \tag{2}$$

を満たすとき，$\hat{\theta}$ は θ の**設計不偏推定量** (design unbiased estimator)，または単に**不偏推定量**と呼ぶ．複雑な標本設計に対して，(2) を厳密に満たす推定量を見つけることは一般に困難である．不偏性のほかに，**漸近設計不偏性** (asymptotic design unbiasedness)

$$E(\hat{\theta} - \theta)/\sqrt{E(\hat{\theta} - \theta)^2} \to 0 \tag{3}$$

や，**漸近設計一致性** (asymptotic design consistency)

$$P\left(\frac{|\hat{\theta} - \theta|}{|\theta|} > \epsilon\right) \to 0, \qquad \text{任意の} \epsilon > 0 \tag{4}$$

などが推定量の良さを計るための重要な基準となる．(3) と (4) における極限は $n, N \to \infty$ のときに計算される．また，条件 $\mathrm{V}(\hat{\theta})/\theta^2 \to 0$ が成立すれば，(3) から (4) を導出できることに注意する．

要素 i が抽出される確率を π_i, 要素 i, j が同時に抽出される確率を π_{ij} で表し，

$$\pi_i = \sum_{s \subset S} P(i \in s), \quad \pi_{ij} = \sum_{s \subset S} P(i, j \in s)$$

π_i を **1 次の包含確率** (first-order inclusion probability) といい，π_{ij} を **2 次の包含確率** (second-order inclusion probability) という．ただし，$\pi_{ii} = \pi_i$ に注意する．非復元確率抽出において，次が成り立つ．

$$\sum_{i=1}^{N} \pi_i = n, \quad \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i$$

ここで，要素 i が抽出されるときに $Z_i = 1$，そうでないときに $Z_i = 0$，という定義関数を考える．すると，Z_i はベルヌーイ確率変数となり，次が成り立つ

$$E(Z_i) = \pi_i, \quad \mathrm{V}(Z_i) = \pi_i(1 - \pi_i). \tag{5}$$

また，Z_i と Z_j の共分散について次が成立する．

$$\mathrm{cov}(Z_i, Z_j) = \pi_{ij} - \pi_i \pi_j \tag{6}$$

実際の標本調査において，**母集団総計値** (population total), $\mathcal{Y} = \sum_{i \in U} y_i$, と**母平均** (population mean), $\mu = \mathcal{Y}/N$, の推定に興味がある場合が多い．ここで y_i は i 番目の要素の特性値を表す．

$\pi_i > 0$ を満たす標本設計 (1) に対して，次の **Horvitz-Thompson 推定量** (Horvitz-Thompson estimator), または π **推定量** (π-estimator)

$$\hat{\mathcal{Y}}_{HT} \equiv \sum_{i=1}^{n} y_i/\pi_i \tag{7}$$

は総計値 $\mathcal{Y}$ の不偏推定量となることが, $\hat{\mathcal{Y}}_{HT} = \sum_{i=1}^{N} Z_i y_i/\pi_i$ と表現し直せば，(5) より簡単に確かめられる．$\hat{\mathcal{Y}}_{HT}$ は加重平均の形をしており，重み $1/\pi_i$ を**設計重み** (design weight) と呼ばれる．さらに，(6) を利用すれば，$\hat{\mathcal{Y}}_{HT}$ の分散も

$$\mathrm{V}\left(\hat{\mathcal{Y}}_{HT}\right) = \sum_{i,j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \tag{8}$$

となることが明快に導かれる．

ところで，通常の標本調査において，１つの要素に対して２つ以上の特徴が同時に調査されるのが一般的である．いま要素 i に対して関心のある特徴 y_i の他に**補助情報** (auxiliary information) である特徴 x_i のデータも記録されているとする．さらに，x についての母集団総計値 $\mathcal{X}$ を既知とする．このとき，実際の調査において，次の**比推定量** (ratio estimator)

$$\hat{\mathcal{Y}}_R \equiv= \frac{\sum_{i=1}^{n} y_i/\pi_i}{\sum_{i=1}^{n} x_i/\pi_i} \mathcal{X} \tag{9}$$

を使うことが多い．

$x_i = 1$ のとき，$\mathcal{X} = N$ となり，比推定量 $\hat{\mathcal{Y}}_R$ は次の **Hájek 推定量** (Hájek estimator)

$$\hat{\mathcal{Y}}_{Haj} \equiv N \frac{\sum_{i=1}^{n} y_i/\pi_i}{\sum_{i=1}^{n} 1/\pi_i} \tag{10}$$

となる．母集団が比較的**均質** (homogeneous) で，包含確率 π_i にばらつきがあり，さらに π_i と y_i の間の相関が強くないとき，Hájek 推定量が Horvitz-Thompson 推定量 (7) よりよい性質を持つ．

3　単純無作為抽出

母集団のすべての要素が同じ確率で抽出される標本設計を，**単純無作為抽出** (simple random sampling) と呼ぶ．この節では，すべての確率抽出法の出発点となる非復元単純無作為抽出法について概説する．このとき，$p(s) = 1/{}_N C_n$ となる．

実際の抽出では，まず母集団のすべての構成要素が抽出される確率が同じとなるように，１つの要素を抽出する．次に，抽出された要素を母集団に戻さず，同じようにしてもう 1 つの要素を抽出する．n 個の要素が抽出されるまでこれを繰り返す．このように，単純無作為抽出法は，母集団の**枠** (frame) と呼ばれる構成要素のリストさえあれば，極めて簡単に実行できる．しかし，母集団が非常に大きいときには枠の作成が非現実的となる場合が多く，単純無作為抽出法に代わる方法を用いる必要がある．また，母集団についての補助情報（たとえば性別の情報）があれば，これを積極的に利用する標本抽出法が望ましいことは言うまでもない．

非復元単純無作為抽出において，1 次と 2 次の包含確率は，それぞれ

$$\pi_i = \frac{n}{N}, \qquad \pi_{ij} = \frac{n}{N} \frac{n-1}{N-1} \tag{11}$$

となる ($i \neq j$). (11) を (7), (10) に代入すると，$\hat{\mathcal{Y}}_{HT} = \hat{\mathcal{Y}}_{Haj}$ となり，π 推定量は，**標本平均** (sample mean), $\bar{y} = n^{-1}\sum_s y_i$, の N 倍

$$\hat{\mathcal{Y}}^{sr} = \sum_{i=1}^{n} \frac{N}{n} y_i = N\,\bar{y} \tag{12}$$

という自然な推定量となる．また (8), (11) より，$\hat{\mathcal{Y}}^{sr}$ の分散は次のようになる．

$$\mathrm{V}\,(\hat{\mathcal{Y}}^{sr}) = N^2(1-f)\frac{\sigma^2}{n} \tag{13}$$

ここで，σ^2 は**母分散** (population variance) である．

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{14}$$

(13) 式における係数 $1-f$ は**有限母集団修正** (finite population correction) と呼ばれる量であり，抽出率が小さいときにはその影響を無視できよう．

母平均 μ の推定に対して，上の議論はほぼそのまま成立する．すなわち，標本平均 $\bar{y}$ は μ の不偏推定量で，その分散は次に与えられる．

$$\mathrm{V}\,(\bar{y}) = (1-f)\frac{\sigma^2}{n} \tag{15}$$

もし，母分散 σ^2 を推定する必要があれば，標本分散

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

を利用することができる．(13) 式の証明と同様にして，$\hat{\sigma}^2$ は σ^2 の不偏推定量，すなわち

$$E\,(\hat{\sigma}^2) = \sigma^2 \tag{16}$$

となることが確かめられる．

ところで，f が小さく，$n, N, N-n$ が共に大きいとき，$\bar{y}$ は近似的に平均 μ, 分散 (15) の正規分布に従うことが知られている．このことと (16) より，μ の信頼係数 $1-2\alpha$ の近似信頼区間は次式で与えられる．

$$P\left(\mu \in \bar{y} \pm z_\alpha\sqrt{1-f}\,\frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 1-2\alpha \tag{17}$$

ただし，z_α は標準正規分布の 100α パーセント点である．

4 層化抽出

大きさ N の母集団 U を，**層** (stratum) と呼ばれる部分母集団 $U_1, \cdots, U_L$ に分割し，各層から独立に確率抽出を行う方法を**層化抽出** (stratified sampling) という．各層における標

本抽出が非復元単純無作為抽出のとき，**層化無作為抽出** (stratified random sampling) という．また母集団の層による分割を**層化** (stratification) という．層化抽出が必要とされる主な理由として，部分母集団の推定の必要性のほか，調査の利便性や（たとえば地方調査部門の活用），推定精度の向上などが考えられる．ここでは，母集団 U の総計値 $\mathcal{Y} = \sum_h \mathcal{Y}_h$ の推定のみについて概説する．ただし，$\mathcal{Y}_h$ は層 U_h の総計値を表す．

まず大きさ N_h の層 U_h から，独立に大きさ n_h の標本 $s_h\,(h = 1, \cdots, L)$ が，一般の確率抽出法によって抽出される場合を考える．標本 s_h の独立性により，π 推定量

$$\hat{\mathcal{Y}}_{st} = \sum_h \hat{\mathcal{Y}}^h_{HT} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}/\pi^h_i \tag{18}$$

は総計値 $\mathcal{Y}$ の不偏推定量となることが分かる．ただし，y_{hi} と π^h_i は層 U_h の要素 i の計量と 1 次の包含確率を表す．また $\mathrm{V}\,(\hat{\mathcal{Y}}_{st}) = \sum_h \mathrm{V}\,(\hat{\mathcal{Y}}^h_{HT})$ により

$$\mathrm{V}\,(\hat{\mathcal{Y}}_{st}) = \sum_{h=1}^{L} \sum_{i,j=1}^{n_h} \frac{\pi^h_{ij} - \pi^h_i \pi^h_j}{\pi^h_i \pi^h_j}\, y_{hi}\, y_{hj} \tag{19}$$

となる．ただし，π^h_{ij} は層 U_h の要素 i, j の 2 次の包含確率を表す．

次に層化無作為抽出の場合を考える．このとき，$f_h = n_h/N_h$ を h 層の抽出率とすると，$\pi^h_i = f_h$ となり，(18) は

$$\hat{\mathcal{Y}}^{sr}_{st} = \sum_h N_h \bar{y}_h \tag{20}$$

となる．ただし，$\bar{y}_h$ は h 層の標本平均を表す．一方，$\mu_h = \mathcal{Y}_h/N_h$ を h 層の平均，$\sigma^2_h = \sum_i (y_{hi} - \mu_h)^2/(N_h - 1)$ を h 層の分散とすると，(19) より $\hat{\mathcal{Y}}^{sr}_{st}$ の分散は次のようになる．

$$\mathrm{V}\,(\hat{\mathcal{Y}}^{sr}_{st}) = \sum_h N^2_h (1 - f_h) \frac{\sigma^2_h}{n_h} \tag{21}$$

推定量 (20) を使うために，各層への**標本配分** (sample allocation), $n_1, \cdots, n_L$，を決める必要がある．最も基本的な標本配分法は，各層の抽出率を一定とする**比例配分** (proportional allocation) である．このとき，$f_h = n_h/N_h = n/N = f$ となる．比例配分のとき，推定量 $\hat{\mathcal{Y}}^{sr}_{st}$ の分散は次となる．

$$V_{prop} = \frac{1 - f}{n} \sum_h w_h \sigma^2_h \tag{22}$$

ただし，$w_h = N_h/N$ は**層重み** (stratum weight) を表す．

最もよい配分法を導出するために，次の**線形費用関数** (linear cost function)

$$C = c_0 + \sum_h c_h\, n_h \quad (c_0 \geq 0, c_h > 0 : \text{既知})$$

を考えるのが便利である．C を固定したとき，(21) が最小になる $n_1, \cdots, n_L$ を**最適配分** (optimum allocation) という．(21) に対してコーシー=シュワルツの不等式を適用すると，最適配分は

$$n_h \propto N_h \sigma_h / \sqrt{c_h}, \quad h = 1, \cdots, L \tag{23}$$

と導出できる．最適配分を行うために，各層の分散 σ_h^2 の情報が必要であることに注意する．最適配分法は，N_h, σ_h^2 が大きい，また単位コスト c_h が安い層に対して，標本サイズ n_h を大きくする方法である．

ところで，各層における単位コスト c_h が同じであると仮定できる場合には，線形費用が標本総数 n に比例する．$c_h = c$ を (23) に代入すると，

$$n_h = n \frac{N_h \sigma_h}{\sum_h N_h \sigma_h}, \quad h = 1, \cdots, L \tag{24}$$

を導出できる．(24) を**ネイマン配分** (Neyman allocation) という．このとき，推定量 $\hat{\mathcal{Y}}_{st}^{sr}$ の分散 (21) は次のようになる．

$$V_{opt} = \frac{1}{n} \left(\sum_h w_h \sigma_h^2 \right)^2 - \frac{1}{N} \sum_h w_h \sigma_h^2 \tag{25}$$

さて，層別の効果について考えよう．各 $1/N_h$ が無視できるとすると，母分散が次のように分解される．

$$\sigma^2 = \sum_h w_h \sigma_h^2 + \sum_h w_h (\mu_h - \mu)^2$$

これを単純無作為抽出における $\hat{\mathcal{Y}}^{sr}$ の分散 (13) に代入すると

$$\mathrm{V}(\hat{\mathcal{Y}}^{sr}) = V_{prop} + \frac{1-f}{n} \sum_h w_h (\mu_h - \mu)^2 \tag{26}$$

となり，比例抽出による推定量の分散が小さくなっていることが分かる．(26) より，特に各層の平均 μ_h がばらつくほど層別の効果が顕著に現れる．一方，(22) と (25) を比較すると，次の関係が成立する．

$$V_{prop} = V_{opt} + \frac{1}{n} \sum_h w_h (\sigma_h - \bar{\sigma})^2 \tag{27}$$

ただし，$\bar{\sigma} = \sum_h w_h \sigma_h$ は各層の標準偏差の加重平均を表す．したがって，σ_h がばらつくほど，ネイマン配分が比例配分より優れていることとなる．以上を纏めると

$$\mathrm{V}(\hat{\mathcal{Y}}^{sr}) \geq V_{prop} \geq V_{opt}$$

となり，層別による推定の精度が改善されていることが分かる．

5 比推定量

目的変数 y と関連する補助変数 x があるとき，x を利用して y の総計値 $\mathcal{Y}$ に対する精度の高い推定量を構築することが可能である．適切なパラメトリックモデルを立てれば，尤度推論などを展開することが望ましいかもしれない（実際これが今日の標本調査の理論の主流である）．しかし，大規模調査において，適切なモデルの選択は一般に難しく，また

最尤法などによる反復計算が必要なため，実際の調査では推定量の使い易さが重要視される．比較的単純でよい性質をもつものに**比推定量** (ratio estimator) がある．

いま x の総計値 $\mathcal{X}$ を既知とする．$\mathcal{X}$ と $\mathcal{Y}$ の π 推定量をそれぞれ $\hat{\mathcal{X}}_{HT}$ と $\hat{\mathcal{Y}}_{HT}$ とすると，$\mathcal{Y}$ の**比推定量** (ratio estimator) は

$$\hat{\mathcal{Y}}_R \equiv \frac{\hat{\mathcal{Y}}_{HT}}{\hat{\mathcal{X}}_{HT}}\mathcal{X} = \frac{\sum_{i=1}^n y_i/\pi_i}{\sum_{i=1}^n x_i/\pi_i}\mathcal{X} \tag{28}$$

と定義される．

5.1 単純無作為抽出

単純無作為抽出のとき，(28) は

$$\hat{\mathcal{Y}}_R = \frac{\bar{y}}{\bar{x}}\mathcal{X} \tag{29}$$

となる．x_i の大きさに応じた標本抽出を行えば，$\hat{\mathcal{Y}}_R$ は不偏推定量になることがある．しかし，単純無作為抽出のもとでは，$\hat{\mathcal{Y}}_R$ は不偏推定量ではなく，その**偏り** (bias) は，n が大きいとき，次のようになる．

$$E(\hat{\mathcal{Y}}_R) - \mathcal{Y} = \frac{N^2(1-f)}{n\mathcal{X}}(R\sigma_x^2 - \sigma_{xy}) \tag{30}$$

ただし，$\sigma_{xy} = \sum_i (y_i - \mu_y)(x_i - \mu_x)/(N-1)$ で, $R = \mathcal{Y}/\mathcal{X}$ である．$\hat{\mathcal{Y}}_R$ の分散は近似的に

$$\mathrm{V}(\hat{\mathcal{Y}}_R) = \frac{N^2(1-f)}{n}(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy}) \tag{31}$$

と評価できる．(31) から, 次の条件

$$\rho > \frac{1}{2}\frac{\sigma_x/\mu_x}{\sigma_y/\mu_y}, \quad \text{ただし } \rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

が成立すれば，$\mathrm{V}(\hat{\mathcal{Y}}_R) < \mathrm{V}(N\bar{y})$ が成り立ち，比推定量 $\hat{\mathcal{Y}}_R$ は通常の推定量 $N\bar{y}$ より優れていることになる．

さらに，有限母集団が次の線形モデル

$$y_i = \beta x_i + \epsilon_i, \qquad i = 1, \cdots, N \tag{32}$$

によって得られているとみなせば，比推定量 $\hat{\mathcal{Y}}_R$ は**最良線形不偏推定量** (BLUE; best linear unbiased estimator) となる．すなわち，$\hat{\mathcal{Y}}_R$ は線形不偏推定量 $\sum_{i=1}^n \ell_i\, y_i$ の中で，最も小さい分散をもつ．ただし，(32) において, x_i を既知とし，確率変数 ϵ_i, ϵ_j は独立で，また $E(\epsilon_i) = 0, V(\epsilon_i) \propto x_i$ を仮定する．また (32) より，$\mathcal{Y} = \beta\mathcal{X} + \sum_{i=1}^N \epsilon_i$ は確率変数なので，$\hat{\mathcal{Y}}_R$ の不偏性は，

$$E(\hat{\mathcal{Y}}_R) = E(\mathcal{Y}) = \beta\mathcal{X}$$

を意味することに注意する．この不偏性を**モデル不偏性** (model unbiasedness) と呼ぶ．

上の議論から分かることは，もし (x_i, y_i) の散布図が大体原点を通る直線上にあり，また y_i の分散が x_i に比例して増大するのであれば，比推定量 $\hat{\mathcal{Y}}_R$ を使うべきであろう．最後に，単純無作為抽出に対しても，次の比型推定量

$$\hat{\mathcal{Y}}_{Rt} = \bar{r}\mathcal{X} + \frac{n(N-1)}{(n-1)\mathcal{X}}(\bar{y} - \bar{r}\bar{x})$$

が設計不偏推定量になっていることを指摘しておこう．ただし，$\bar{r} = n^{-1}\sum_{i=1}^{n} y_i/x_i$ である．

5.2 層化無作為抽出

まず各層の総計値 $\mathcal{X}_h$ を既知とし，比推定が各層に適用して得られる**個別比推定量** (separate ratio estimator)

$$\hat{\mathcal{Y}}_{Rs} = \sum_h \frac{\bar{y}_h}{\bar{x}_h}\mathcal{X}_h \tag{33}$$

が考えられる．一方，**結合比推定量** (combined ratio estimator)

$$\hat{\mathcal{Y}}_{Rc} = \frac{\sum_h N_h \bar{y}_h}{\sum_h N_h \bar{x}_h}\mathcal{X} \tag{34}$$

においては，総計値 $\mathcal{X}$ のみが必要とされる．一般に $\hat{\mathcal{Y}}_{Rs}$ の偏りが $\hat{\mathcal{Y}}_{Rc}$ のそれに比べて大きく，特に n_h が大きくないときに注意が必要であろう．

各 n_h が比較的大きいとき，$\hat{\mathcal{Y}}_{Rs}$ と $\hat{\mathcal{Y}}_{Rc}$ の分散は

$$\begin{aligned}
\mathrm{V}(\hat{\mathcal{Y}}_{Rs}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h}(\sigma_{yh}^2 + R_h^2\sigma_{xh}^2 - 2R_h\sigma_{xyh}) \\
\mathrm{V}(\hat{\mathcal{Y}}_{Rc}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h}(\sigma_{yh}^2 + R\sigma_{xh}^2 - 2R\sigma_{xyh})
\end{aligned}$$

となり，特に各層の母平均の比 R_h がばらつくとき，$\mathrm{V}(\hat{\mathcal{Y}}_{Rs}) < \mathrm{V}(\hat{\mathcal{Y}}_{Rc})$ が成り立つ．どの推定量を使うべきかは**平均 2 乗誤差** (mean squared error) などの基準で総合的に判断する必要があろう．

上の議論は，n_h が既知のときの話であって，n を固定して，$\mathrm{V}(\hat{\mathcal{Y}}_{Rs})$ を最小にするように，標本配分を求めることも可能である．このときの最適配分は

$$n_h = n\frac{N_h\sigma_{dh}}{\sum_h N_h\sigma_{dh}}, \quad h = 1, \cdots, L$$

となる．ただし，σ_{dh}^2 は次式で定義される．

$$\sigma_{dh}^2 = (N_h - 1)^{-1}\sum_{i=1}^{N_h}(y_{hi} - R_h x_{hi})^2$$

6　標本誤差のジャックナイフ推定

層化抽出における比推定量に代表されるように，複雑な標本設計に基づく推定量の偏りや分散などの**標本誤差** (sampling error) の評価は一般に容易ではない．これまでに示した標本誤差の近似評価は主として**デルタ法** (delta method) と呼ばれる方法を用いて行われた．しかし，より複雑な場合においてデルタ法に限界がある．そのため，近年，複雑な場合にも適用できる，**ブートストラップ法** (bootstrap method) や**ジャックナイフ法** (jackknife method) などの計算機指向型手法がよく利用される．ここで，計算時間をそれほど必要としないジャックナイフ法による誤差の推定について概説する．

まず非層化確率抽出法により大きさ n の標本 s が得られた場合を考える．標本 s に基づく母数 θ に対する推定量を $\hat{\theta}$ とし，i 番目のデータを取り除いたときの推定量を $\hat{\theta}_{i\prime}$とする．**ジャックナイフ擬似値** (jackknife pseudovalue) を，復元抽出のときに

$$\tilde{\theta}_{i\prime} = n\hat{\theta} - (n-1)\hat{\theta}_{i\prime} \tag{35}$$

と定義し，非復元抽出のときには

$$\tilde{\theta}_{i\prime} = n\hat{\theta} - (n-1)\left\{\hat{\theta} - \sqrt{1-f}\,(\hat{\theta} - \hat{\theta}_{i\prime})\right\} \tag{36}$$

と定義する．ただし，$f = n/N$ は抽出率で，$f \approx 0$ のとき，(36) と (35) の差は無視できよう．

すると，$\hat{\theta}$ より偏りの少ない**ジャックナイフ推定量** (jackknife estimator) を

$$J(\hat{\theta}) = n^{-1}\sum_{i=1}^{n}\tilde{\theta}_{i\prime} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \tag{37}$$

と構成できる．ただし，$\hat{\theta}_{(\cdot)} = n^{-1}\sum\hat{\theta}_{i\prime}$である．さらに，$\hat{\theta}$ の分散のジャックナイフ推定量も

$$\begin{aligned} V_J(\hat{\theta}) &= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{\theta}_{i\prime} - J(\hat{\theta}))^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{i\prime} - \hat{\theta}_{(\cdot)})^2 \end{aligned} \tag{38}$$

と構成できる．

次に層化抽出の場合について考える．一般論の展開は難しく，ここではまず推定すべき母数 θ が

$$\theta = \sum_{h=1}^{L}\ell_h\,\theta_h \tag{39}$$

で表現できると仮定する．ただし，$\ell_h > 0$ は既知で，θ_h は h 層の特性値を表す．合計 $\mathcal{Y} = \sum_h \mathcal{Y}_h$ や平均 $\mu = \sum_h w_h\mu_h$ などが (39) で表せることに注意する．次に各層からの標本抽出は独立で，また θ の推定量 $\hat{\theta}$ は

$$\hat{\theta} = \sum_{h=1}^{L}\ell_h\,\hat{\theta}_h \tag{40}$$

と $\hat{\theta}_h$ の線形結合で表せる場合を考える．ただし，$\hat{\theta}_h$ は h 層からの標本に基づく θ_h に対する推定量である．個別比推定量などが (40) の例である．このとき，$\hat{\theta}_h$ の独立性から，次が言える．

$$\mathrm{V}(\hat{\theta}) = \sum_{h=1}^{L} \ell_h^2 \, \mathrm{V}(\hat{\theta}_h)$$

したがって，ジャックナイフ法を各層に対して行えば，$\hat{\theta}$ の分散のジャックナイフ推定量は

$$V_{Js}(\hat{\theta}) = \sum_{h=1}^{L} \ell_h^2 \, V_J(\hat{\theta}_h) \tag{41}$$

と得られる．ただし，$V_J(\hat{\theta}_h)$ は (38) に基づいて計算される．また，各層に対して (37) から $J(\hat{\theta}_h)$ を計算すれば，次のジャックナイフ推定量が得られる．

$$J_s(\hat{\theta}) = \sum_{h=1}^{L} \ell_h \, J(\hat{\theta}_h) \tag{42}$$

参考文献

[1] W. G. Cochran, *Sampling Techniques*, 第 3 版, John Wiley, 1977. 【標本調査の理論と実際の原点とも言うべく，依然として名著である】

[2] C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer, 2003. 【統計学の本流に位置づけられるべく，モデルに基づく標本調査論を中心とした記述がなされている】

[3] 林知己夫 (編), 社会調査ハンドブック, 朝倉書店, 2002. 【ユーザー向けの実用的解説書である】

[4] 竹内啓 (編), 統計学辞典, 東洋経済新報社, pp.242–257, 1989. 【標本調査法の全般についてコンパクトに纏められ，また特に日本語による文献の入門にもなろう】