⑥ 層別とその推定値への影響

青山博次郎

サンプリングに於て，層別を行うことは通常のようになつている。この時分析の段階ではAという性質について層別した各層の統計値（例えば標本平均）を欲しいのであるが，適当な資料が予め得られなかつたためにBという性質で層別してサンプリングを行うことがある。このようにして得られたdataを発表に際してはAという性質で層別して行う場合どれ位の偏りを生ずるかを考えてみることにしよう。

簡単な実例では学級数の各段階毎の標識の標本平均を計算するのに，サンプリングでは資料の関係から生徒数を用いた場合などがこれに当る。

扨て標識をYとし，性質Bについて層別し，これより比例抽出法を用いてサンプルを抽出したとする。このサンプルを更に性質Aについて層別しなおすのであるから二回抽出法の公式が使える。全平均 $\bar{Y}$ の推定値 $\bar{y}$ は

$$\bar{y} = \frac{1}{n}\sum_{j}\sum_{k} y_{jk} = \frac{1}{n}\sum_{j}\sum_{k}\sum_{i} y_{ijk} \qquad (1)$$

但しjは性質Aによる層，iは性質Bによる層を表わし，nはサンプル数である。

このとき

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i}\sum_{k} y_{ik}\right) = \frac{1}{n}\sum_{i}\frac{n_i}{N_i}\sum_{k=1}^{N_i} Y_{ik} = \frac{1}{N}\sum_{i}\sum_{k} Y_{ik} = \bar{Y} \qquad (2)$$

となつて unbiased である.

また分散は

$$V(\bar{y}) = \frac{\sigma^2}{n} - \frac{1}{n}\sum_i P_i(\bar{Y}_i - \bar{Y})^2 \qquad (3)$$

であつてAによる層別の影響はない。(註)

次に各層毎の推定平均について考えると

$$\bar{y}_i = \frac{1}{n_i}\sum_j\sum_k y_{ijk} = \frac{1}{n_i}\sum_k y_{ik} \qquad (4)$$

$$E(\bar{y}_i) = \bar{Y}_i \qquad (5)$$

となるから $\bar{y}_i$ については unbiased なることは勿論であるが

$$\bar{y}_j = \frac{1}{n_j}\sum_i y_{ij} = \frac{1}{n_j}\sum_i\sum_k y_{ijk} \qquad (6)$$

は unbiased とならない。その理由に断るまでもなく分母の n_j が確率変数だからである。

さて一般に二つの確率変数 X, Y について

$$\frac{X}{Y} = \frac{E(X)}{E(Y)} + \frac{XE(Y) - YE(X)}{E^2(Y)} + R, \quad R\text{は誤差項} \qquad (7)$$

が成立つ。そこで

$$E\left(\sum_k\sum_i y_{ijk}\right) = \sum_i n_{ij} \sum_{(n_{ij}\cdots n_{ij''})} P_{ij}^{n_{ij}} P_{ij'}^{n_{ij'}} \cdots P_{ij''}^{n_{ij''}} \frac{n_i!}{n_{ij}!n_{ij'}!\cdots n_{ij''}!} \cdot \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}} Y_{ijk}$$

$$= \sum_i \frac{n_i}{N_i} Y_{ij} = \frac{n}{N} Y_j \qquad (8)$$

但し $P_{ij} = N_{ij}/N_i$ 等, $(n_{ij}\cdots n_{ij''})$ は $n_{ij}+\cdots+n_{ij''} = n_i$ なるときの $n_{ij}, \cdots, n_{ij''}$ のすべての組合せについての和を示す。

(註) 著者：二回抽出法について，統数研講究録，第6巻第6号（昭和25年9月）参照。
記号や条件もこの論文中のものを用いてある。

同様にして

$$E(n_j) = \frac{n}{N} N_j \tag{9}$$

従って (7) により，(8)，(9) を用いて

$$E(\bar{y}_j) \fallingdotseq \frac{E(\sum_k \sum_i y_{ijk})}{E(n_j)} = \frac{Y_j}{N_j} = \bar{Y}_j \tag{10}$$

となるから近似的に *unbiased* である．

次に (7) の $E(R)$ を評価してみると

$$E(R) \fallingdotseq \frac{E(X)}{E^2(Y)} \left(E(Y^2) - E^2(Y) \right) = \frac{E(X) V(Y)}{E^3(Y)} \tag{11}$$

従って同様の計算を行うと

$$V(n_j) = E(n_j^2) - E^2(n_j) = \frac{n}{N} N_j - \frac{n}{N} \sum_i \frac{N_{ij}^2}{N_i} \tag{12}$$

$$\therefore \quad E(R) \fallingdotseq \frac{N}{n} \frac{\bar{Y}_j}{N_j} \left(1 - \sum_i \frac{N_{ij}^2}{N_i N_j} \right) = O\left(\frac{1}{n}\right) \tag{13}$$

即ち

$$E(\bar{y}_j) = \bar{Y}_j + O\left(\frac{1}{n}\right) \tag{14}$$

之が (10) の代りに用いるべき式である。

また $\bar{y}_j$ の分散は

$$V\left(\frac{X}{Y}\right) \fallingdotseq \frac{E(X - \varphi Y)^2}{E^2(Y)} \quad , \text{但し} \quad \varphi = \frac{E(X)}{E(Y)} \tag{15}$$

を用いて計算を行えば y_{ij} と n_{ij} の相関係数 ρ として（通常は正となることが多いから大きい目にみつもって）

$$V(\bar{y}_j) = \frac{1}{N_j^2} \sum_i \frac{N_{ij}^2 \sigma_{ij}^2}{n_{ij}} + \frac{N}{n N_j^2} \sum_i N_{ij} \left(1 - \frac{N_{ij}}{N_i} \right) \left(\bar{Y}_{ij}^2 + \bar{Y}_j^2 + \frac{\sigma_{ij}^2}{m_{ij}} \right) \tag{16}$$

となることが分る。