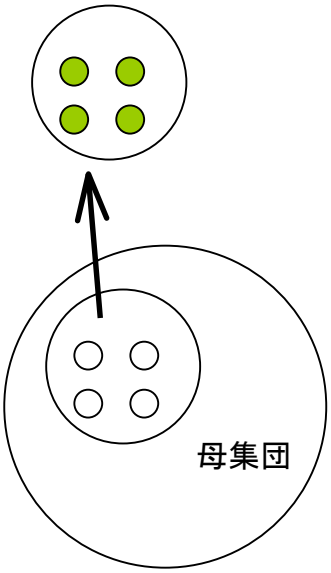
2000/4/24
第　2　講

比較統計システム論　講義録（2000 年度）

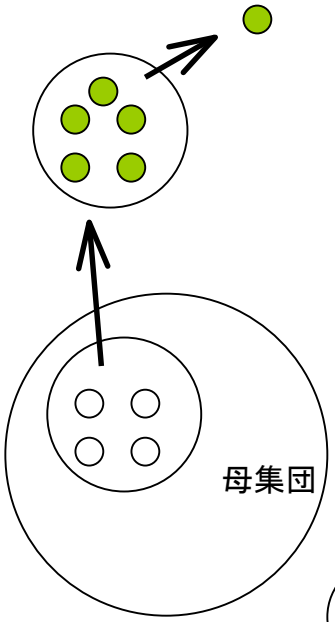
標本抽出法（sampling techniques）について

様々な抽出法のイメージ

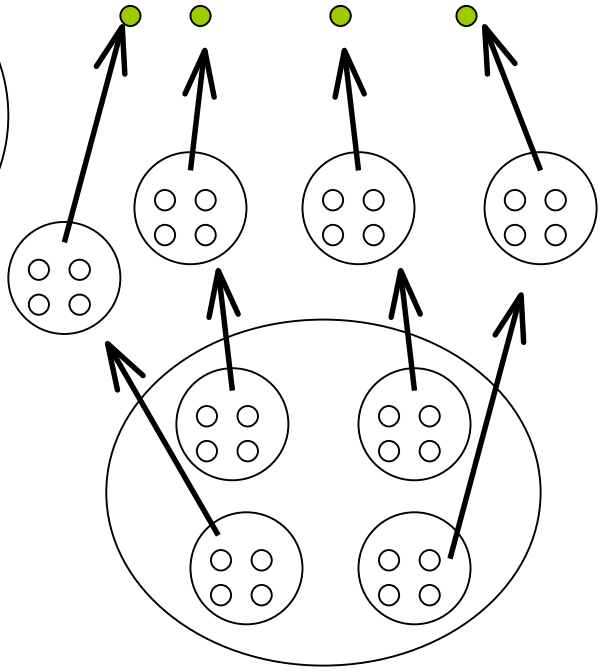
集落抽出法



2段抽出法



層化抽出法



集落抽出法

調査単位を抽出するのではなく、いくつかの調査単位の集まり（集落）を抽出単位として抽出する方法を集落抽出法という。

政府統計では、全国の全世帯を 50 世帯程度の調査区に区分し、調査区を抽出する方法をとっているものがある。

集落抽出法では、できるだけ集落を等質にしておくことが重要である。この方法は、調査に要する負担が軽いなど実務面で優れているが、標本数が同じであれば、通常は単純無作為抽出法より精度は低くなる。

2段抽出法

母集団から標本を抽出するのではなく、まず集落を抽出し、抽出された集落から標本を抽出する方法を2段抽出法という。最初に調査区をいくつか抽出し、さらにそれらの中から必要な標本を抽出する方法である。

層化抽出法

2段抽出において、母集団がいくつかの集落に分割されている状態で各集落から調査単位を抽出するとき、多くの場合、各集落を層（stratum）、集落への分割を層化（stratification）、この抽出方法を層化抽出法（stratified sampling）という。

集落を全く無作為に作るのではなく、それぞれに何らかの特性を持たせることで、母集団の情報をより適切に抽出することができる。

この方法の基本的な考え方は、次のようなものである。すなわち、単純無作為抽出における標本平均の分散は、標本平均という推定量の平均的な誤差であるから、分散が小さい場合には精度が高いと解釈できる。推定式の分散の式を観察すると、標本数を増やせば精度が高くなり、また母分散が小さい場合には、大きい場合よりも精度が高くなることがわかる。

このことは、母分散の大きな母集団の平均値を調べる場合には、多くの標本を必要とし、母分散の小さな母集団の場合には、少ない標本数でも同じ精度を得ていることを意味している。

すなわち、母分散を小さくできれば調査の精度は向上するのだとすれば、推定値の散らばりの小さな層毎に無作為抽出を行って標本平均の分散を小さくし、それらをまとめて母集団全体の平均の推定の精度を向上させようというのが層化抽出法の考え方である。

数学的表現を用いると、大きさ N の母集団から、大きさnの単純ランダムサンプルを抽出するとき、母平均の推定量としての標本平均$\bar{Y}$の分散 V は

$$V = \frac{N-n}{N-1} \bullet \frac{\sigma^2}{n}$$

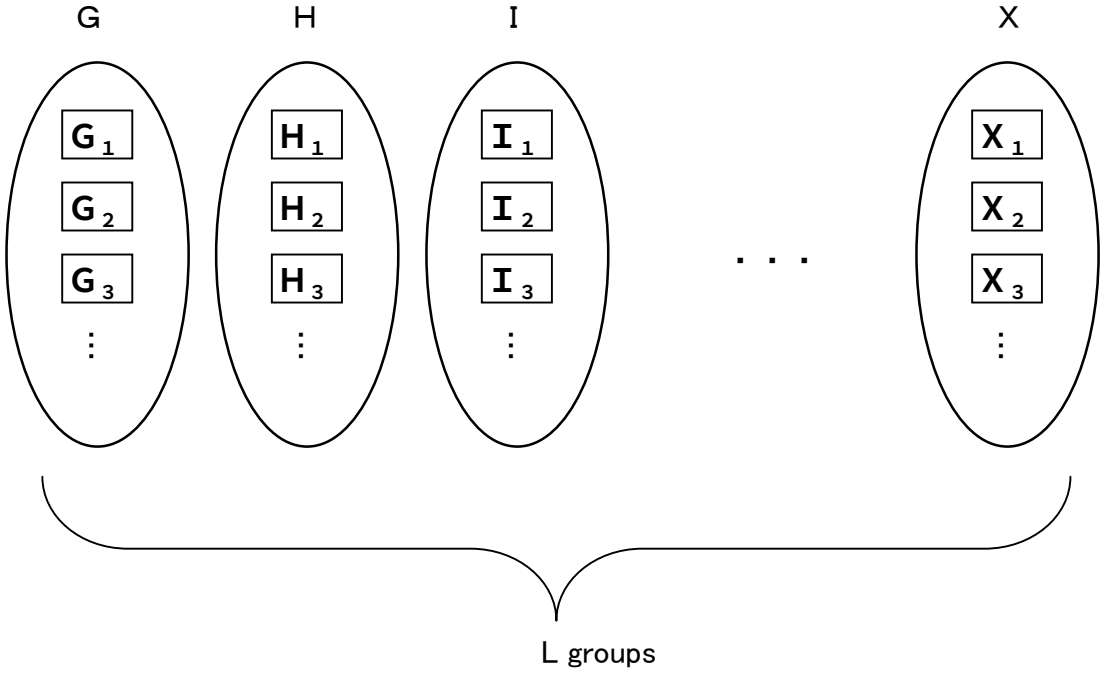
である。N とσ^2を固定すると、nを増やせば V は減少するし、N とnを固定すると、V は母分散σ^2が小さいほど小さい。
nとσ^2を固定すると、

$$\frac{N-n}{N-1} = \frac{(N-1)-(n-1)}{N-1} = 1 - \frac{n-1}{N-1}$$

であり、V は N の増加関数であることがわかる。

層化抽出法を少し厳密に表現すると次のようになる。

単純ランダムサンプリングでは、母集団を構成する各個体は 1 枚のカードであり、表側には番号が書かれており、裏側には項目 y の値が書かれている。層化抽出法では母集団に含まれる各個体が L 個のグループに分類されており、その個体がi番目のグループ(G_i)に属するときは、カードの表側にG_iと書かれている。



ここで標本を抽出するときのルールとして、「標本を抽出する段階では、カードの裏側を見てはいけないが、表側の情報は自由に利用することができる」、あるいは「標本を抽出する段階で利用できる情報は表側に書かれており、標本として抽出された個体についてのみ観測される情報は裏側に書かれている」とする。そこで、母集団から大きさnの標本を次の手順で抽出しよう。

(1)カードをグループごとにまとめる。

（2）各グループの枚数を数える。i番目のグループG_iの枚数をN_iで表わす。
（3）各グループから、それぞれ大きさ N_i（i＝1、2、…L）の単純ランダムサンプルを抽出する。ただし、$N_1＋N_2＋\cdots＋N_L＝N$ である。
この抽出法を層化ランダムサンプリングという。標本抽出理論では、グループを層、グループに分けることを層化、母集団での各グループに属する個体数N_iを第i層の大きさという。

母集団としての第i層の平均を$\bar{y}_i$、分散をσ_iとする。ここで母集団の平均、分散と、層ごとの平均、分散との関係を調べる。第i層の中でのj番目の個体のyの値をy_{ij}とする。
$W_i＝N_i／N$とおき、第i層の重みという。

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{Ni}y_{ij}=\sum_{i=1}^{L}\frac{N_i}{N}\frac{1}{N_i}\sum_{j=1}^{N_i}y_{ij}=\sum_{i=1}^{L}W_i\bar{y}_i$$

母平均$\bar{y}$は、各層の平均$\bar{y}_i$に各層の重み W_iを付けた加重平均で表わせる。

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{Ni}(y_{ij}-\bar{y})^2=\sum_{i=1}^{L}\frac{N_i}{N}\cdot\frac{1}{N_i}\sum_{j=1}^{N_i}\left[(y_{ij}-\bar{y}_i)+(\bar{y}_i-\bar{y})\right]^2$$

$$=\sum_{i=1}^{L}W_i\frac{1}{N_i}\sum_{j=1}^{Ni}\left[(y_{ij}-\bar{y}_i)^2+(\bar{y}_i-\bar{y})^2+2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})\right]$$

$$=\underbrace{\sum_{i=1}^{L}W_i\sigma_i^2}+\underbrace{\sum_{i=1}^{L}W_i(\bar{y}_i-\bar{y})^2}$$

各層の分散の加重平均（層内分散＝σ_w^2）　**各層の平均と母平均の差の自重の加重平均**（層間分散＝σ_b^2）

母分散は層内分散と層間分散の和として表わせる。

層化ランダムサンプルに基づく母平均の推定を考える。第i層からの大きさ N_iの単純ランダムサンプルの標本平均を$\bar{Y}_i$とする。$\bar{Y}_i$からy_iの不偏推定量であることと、上の母平均と層の平均の関係から、

$$\bar{Y}_{st}=\sum_{i=1}^{L}W_i\bar{Y}_i$$

は、母平均の不偏推定量であることがわかる。この$\bar{Y}_{st}$を層化推定量と呼ぶ。添え字のstは層化（stratification）を表わす。$\bar{Y}_{st}$の分散は

$$V(\bar{Y}_{st}) = \sum_{i=1}^{L} W_i^2 V(\bar{Y}_i) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$$

である。

この式の導出には、「L個の標本平均$\bar{Y}_i$（i＝１、２、…、L）は独立である」ということが使われている。これは「各層からの単純ランダムサンプルの抽出は、他の層の抽出とは無関係に行われる」ということによって正当化される。

ここで、層化ランダムサンプルを用いると同じ大きさの単純ランダムサンプルよりも分散の小さな母平均の推定量が得られる可能性について考えてみよう。

最初に、層化は既に定まっているものとして、各層の標本数を定める問題を考えよう。これを、標本数の割り当て問題という。

全体の標本数をnとするとき、層の大きさに比例するように各層の標本数を定める方法を、比例割当という。すなわち、第i層の標本数をn_iとすると

$$n_i = \frac{N_i}{N} n = W_i n$$

である。また分散は、

$$V(\bar{Y}_{st}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$$

$$= \sum_{i=1}^{L} W_i^2 \frac{NW_i - nW_i}{NW_i - 1} \cdot \frac{\sigma_i^2}{nW_i} = \frac{N-n}{n} \sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{NW_i - 1}$$

$$= \frac{N-n}{(N-1)n} \sum_{i=1}^{L} W_i^2 \sigma_i^2 \left(1 + \frac{1 - W_i}{NW_i - 1} \right)$$

となる。$NW_i = N_i$がある程度大きければ、（１－W_i）／（NW_i－１）は無視して、

$$V(\bar{Y}_{st}) \approx \frac{N-n}{(N-1)n} \sum_{i=1}^{L} W_i \ \sigma_i^2$$

と近似できる。

層内分散（within-variance）をσ_w^2、層間分散（between-variance）をσ_b^2で表わすと、

$$V(\bar{Y}_{st}) \approx \frac{N-n}{N-1} \cdot \frac{\sigma_w^2}{n}$$

となる。大きさnの単純ランダムサンプルの標本平均$\bar{Y}$の分散は、母分散が層内分散と層間分散の和として表わせることを利用して、

$$V(\bar{Y}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = \frac{N-n}{N-1} \cdot \frac{\sigma_w^2 + \sigma_b^2}{n}$$

これより

となる。これは、比例割当を用いた場合の層別推定量の分散は、同じ大きさの単純ランダムサンプルの標本平均の分散より小さく、比で考えるとσ_W^2／σ^2倍になるといえる。

$$\frac{V(\overline{Y}_{st})}{V(\overline{Y})} \approx \frac{\sigma_w^2}{\sigma_w^2+\sigma_b^2}$$

逆に、層別推定量の分散を最小にするような割り当てを考えることも重要である。これをネイマン割当という。

一般的な層別推定量の分散は

$$\sum_{i=1}^{L} W_i(\overline{Y}_{st}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$$

である。

$$V(\overline{Y}_{st}) = \sum_{i=1}^{L} W_i^2 \frac{N_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i} - \sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{N_i - 1}$$

という条件の下で、この$V(\overline{Y}_{st})$を最小にするn_1、…n_Lを求める。

その際、ラグランジェ乗数法を用いる。

$$\frac{\partial V}{\partial n_i} = -W_i^2 \frac{N_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i^2}$$

$\sum_{i=1}^{L} n_i$ をn_iで偏微分すると

（A）
$$-W_i^2 \frac{N_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i^2} + \lambda = 0$$

をn_iで偏微分すると1であり、未定乗数をλとすれば

（B）
$$n_i = \sqrt{\frac{1}{\lambda}} \sqrt{\frac{N_i}{N_i - 1}} W_i \sigma_i \qquad (i=1, \cdots, L)$$
$$\sum_{i=1}^{L} n_i = n$$

を得る。（B）をn_iについて解くと

$$n_i = \frac{\sqrt{\frac{N_i}{N_i - 1}} W_i \sigma_i}{\sum_{i=1}^{L} \sqrt{\frac{N_i}{N_i - 1}} W_i \sigma_i} n$$

となり、

ここでも、N_iがある程度大きいとして、N_i／（N_i－1）を1とみなせば、

(i=1, … , L)

(C)

$$\sigma_i = \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2 \qquad \sum_{i=1}^{N}(y_i-\bar{y})^2$$

こうして求められる層化推定量の分散を最小にする割合をネイマン割当てという。

(C)より、ネイマン割当てでは、各層の標本数は層の大きさと層の標準偏差の積に比例している。層の分散が同じなら、個体数の多い層に、層の個体数が同じなら分散の大きい層により多く標本数を割り当てることを意味している。

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

母分数 であったが、 をNで割る代わりにN－1で割ったもの $\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2$ $\sigma^2 = \frac{N-1}{N}\tilde{\sigma}^2$ のを σ^2 で表わす。

本来の母分散との関係は

$$\tilde{\sigma}^2 = \frac{1}{N_i-1}\sum_{i=1}^{N}(y_{ij}-\bar{y}_i)^2$$

あるいは

$V(\bar{y}) = \frac{N-n}{N}\frac{\tilde{\sigma}^2}{n} = (1-f)\frac{\tilde{\sigma}^2}{n} = \frac{\tilde{\sigma}^2}{n} - \frac{\tilde{\sigma}^2}{N}$

である。$\tilde{\sigma}^2$ は修正母分散と呼ぼう。Nが大きければ σ^2 と $\tilde{\sigma}^2$ の違いは無視できる。

各層に対しても修正母分散を考えると、

(i=1,…, L)

で表わす。 $\frac{W_i\tilde{\sigma}_i}{\sum_{i=1}^{L}W_j\tilde{\sigma}_j}n$

ここでf＝n／N（抽出率）である。

ネイマン割当は、第i層の標本数n_iに対して、

$$V(\bar{Y}_{st}) = \frac{(\sum_{i=1}^{L}W_i\tilde{\sigma}_i)^2}{n} - \frac{\sum_{i=1}^{L}W_i\tilde{\sigma}_i^2}{N}$$

と書ける。

ネイマン割当の場合の層化推定量の分散は

$$n_i \approx \frac{W_i\ \sigma_i}{\sum_{i=1}^{L}W_i\ \sigma_i}n$$

となる。

単純ランダムサンプルの標本平均、比例割当による層化推定量、ネイマン割当による層化推定量の分散（標本サイズはn_iを、V_{ran}、V_{prop}、V_{ney}と表わし、修正分散を用いると、

$$V_{prop} = (1-f)\frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}{n} = \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}{n} - \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}{N}$$

$$V_{ney} = \frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{n} - \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}{N}$$

$$= \frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{n} - \frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{N} - \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 - \left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{N}$$

$$= (1-f)\frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{n} - \frac{\sum_{i=1}^{L} W_i (\tilde{\sigma}_i - \overline{\tilde{\sigma}})^2}{N}$$

$$\overline{\tilde{\sigma}} = \sum_{i=1}^{L} W_i \tilde{\sigma}_i$$

$$\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 - \left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 = \sum_{i=1}^{L} W_i (\tilde{\sigma}_i - \overline{\tilde{\sigma}_i})^2$$

ただし、 $\tilde{\sigma}^2 = \sum_{i=1}^{L} \frac{N_i - 1}{N-1}\tilde{\sigma}_i^2 + \sum_{i=1}^{L} \frac{N_i}{N-1}(\overline{y_i} - \overline{y})^2$ である。

ここで、

$$V_{ran} = V_{prop} + \frac{N-n}{n(N-1)}\left[\sum_{i=1}^{L} W_i (\overline{y_i} - \overline{y})^2 - \frac{1}{N}\sum_{i=1}^{L}(1-W_i)\tilde{\sigma}_i^2\right]$$

$$V_{prop} = V_{ran} + \frac{1}{n}\sum_{i=1}^{L} W_i (\tilde{\sigma}_i - \overline{\tilde{\sigma}})^2$$

を用いると、

$$\sum_{i=1}^{L} W_i (\tilde{\sigma}_i - \overline{\tilde{\sigma}})^2$$

を得る。$\tilde{\sigma}_{ran}^2 = (1-f)\frac{\tilde{\sigma}^2}{n} = \frac{\tilde{\sigma}^2}{n} - \frac{\tilde{\sigma}^2}{N}$

$\overline{\tilde{\sigma}}$はL個の層の標準偏差の分散である。$V_{ney}$は$V_{prop}$以下であり、かつ各層

の標準偏差の違いが大きいほど減少量が大きいことを示している。

参考文献

Cochran, William, G. (1977) Sampling Techniques, 3rd ed, NeW York: John Wiley & Sons.

Kish, Leslie. (1965) Survey Sampling, NeW York: John Wiley &Sons.

Levy, Paul, S. and LemeshoW, Stanley. (1999) Sampling of Population, 3rd ed, NeW York: John Wiley & Sons.

豊田秀樹(1998)『調査法講義』、朝倉書店。

鈴木達三・高橋宏一(1998)『標本調査法』、朝倉書店。

杉山明子(1984)『社会調査の基本』、朝倉書店。

西平重喜(1985)『統計調査法 改訂版』、培風館。