有意抽出時の標本誤差とサンプルサイズ、母分散の関係

太郎丸博 （京都大学）

1 問題

サンプリング理論は無作為抽出した標本を前提に作られており、有意抽出した標本（以下、有意標本と略称）の性質についてはざっと検索した限り研究が見当たらない。しかし、多くの学問分野で有為標本は活用されており、その性質について理解することは、それらの研究成果を理解する上で重要である。この報告では、有意標本からある変数 Y の母集団における平均値（以下、母平均と略称）を推測する際に、標本サイズと母分散が標本誤差（次節で定義）にどのように影響するのか論じる。

この研究では以下の 3 つを主張する。

1. 有意標本でも標本誤差は、母集団での標準偏差（以下、母標準偏差と略称）に反比例する。
2. 有意抽出された標本（以下、有意標本）でも、標本誤差は、サンプルサイズの平方根に反比例する。
3. 有意抽出した標本の標本誤差は、無作為抽出した場合の標本誤差より大きくなる。

1 については次節で証明する。2 と 3 については一般的な答えを出す用意がないので、ある特定の条件で作り出した 50 個の有意標本を例に検討する。2 については有意標本でも independently and identically distributed (iid) という条件をつければ証明できることを補論で示す。

2 標本誤差と母標準偏差の関係

Y という変数の母平均 $\mathrm{m}(Y)$ を推測しようとしているとする。母集団から N 個の事例 i $(i = 1, \ldots, N)$ を有意抽出し、標本とする。これを M 回繰り返すとする。j 番目の標本の Y の平均を $\mathrm{m}(Y_j)$ $(j = 1, \ldots, M)$ とする。j番目の標本のi番目の事例のYの値をY_{ij}とすると、

$$\mathrm{m}(Y_j) = \frac{\sum_{i=1}^{N} Y_{ij}}{N} \qquad (1)$$

と定義する。また、標本誤差 $\mathrm{s}(Y)$ は、

$$\mathrm{s}(Y) = \sqrt{\frac{\sum_{j=1}^{M}\left(\mathrm{m}(Y_j) - \mathrm{m}(Y)\right)^2}{M-1}} \qquad (2)$$

と定義する。ここで証明したいのは以下の定理である。

定理 1: $\mathrm{s}(aY) = a \times \mathrm{s}(Y)$

証明

Y を a 倍すると、母平均も a 倍になる（証明省略、統計の教科書をあたれ)。すなわち、

$$\mathrm{m}(aY) = a \times \mathrm{m}(Y) \qquad (3)$$

である。同様に有意標本であっても Y_{ij} を a 倍すると、以下のように標本平均もa倍になる。

$$\mathrm{m}(aY_j) = \frac{\sum_{i=1}^{N} a\,Y_{ij}}{N} = a \times \frac{\sum_{i=1}^{N} Y_{ij}}{N} = a \times \mathrm{m}(Y_j) \qquad (4)$$

それゆえ、以下のように Y をa倍すると、標本誤差も a 倍になる。

$$\begin{aligned} \mathrm{s}(aY) &= \sqrt{\frac{\sum_{j=1}^{M}\left(\mathrm{m}(aY_j) - \mathrm{m}(aY)\right)^2}{M-1}} = \sqrt{\frac{\sum_{j=1}^{M}\left(a \times \mathrm{m}(Y_j) - a \times \mathrm{m}(Y)\right)^2}{M-1}} \\ &= \sqrt{\frac{a^2 \sum_{j=1}^{M}\left(\mathrm{m}(Y_j) - \mathrm{m}(Y)\right)^2}{M-1}} = a\sqrt{\frac{\sum_{j=1}^{M}\left(\mathrm{m}(Y_j) - \mathrm{m}(Y)\right)^2}{M-1}} = a \times \mathrm{s}(Y) \qquad (5) \end{aligned}$$

以上で証明終わり。

3 標本誤差とサンプルサイズの関係

データと方法

2021 年 5 月 23〜24 日に、私の講義を受講している学生にオンラインで以下のような有意抽出をするように指示した。「1〜100 の数字の中から 10 個好きな数字を自由に選んで下さい（同じ数字を２回以上選ばないでください)」。その結果 50 人から回答を得た。これらの数字を母集団における事例の ID とみなすと、$M = 50,\ N = 10$ である。理解を容易にするために、50 人の学生が科学社会学の研究のために、ある分野で出版された 100 本の論文の中から、それぞれ 10 本ずつ有意抽出し、被引用回数の母平均を推測しようとしていると考える。各 ID の論文の被引用回数は、母平均＝21、母標準偏差＝22 で、ID 番号の二次関数に近似するように設定した。

実際には一人あたり 10 本抽出しただけだが、その 10 個の中から n 個のサンプルを無作為抽出（非復元リサンプリング）し、n個のサンプルを 50 回有意抽出したとみなす。 このn個のサンプルに関して標本誤差を計算する。nを 2 から 10 まで変化させたときの標本誤差の大きさを計算する。以上を 500 回繰り返して、標本誤差と標本サイズの関係を検討する。

最後に、同じ母集団から無作為標本を有意標本と同様に$M = 50,\ N = 10$ だけ抽出し、その際の標本誤差を計算する。これを 500 回繰り返して、その結果を有意標本の標本誤差と比較する。

シミュレーションの結果

被引用回数の標本平均を 50 個の標本についてすべて計算し、それらのヒストグラムを描いたのが、図 1 左パネルである。母平均よりもやや高めになっている。

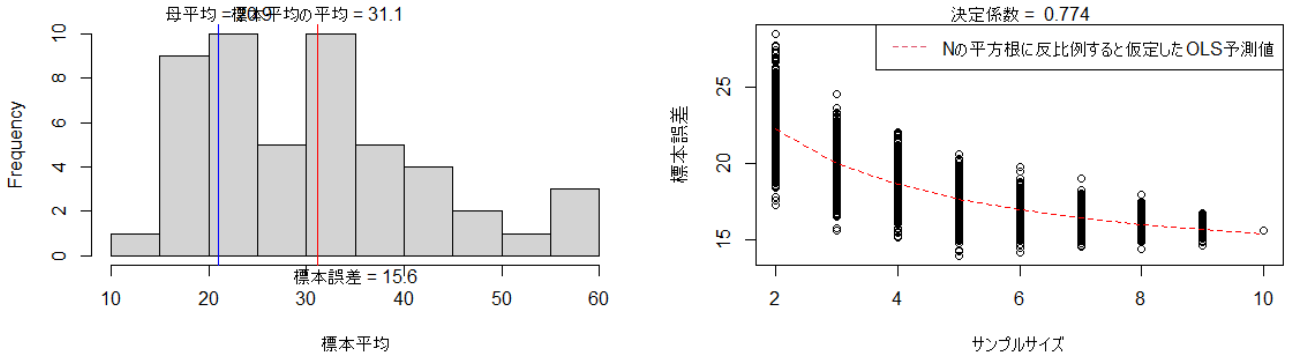

図 1 標本平均のヒストグラム（左パネル）と有意標本のサイズと標本誤差の関係（右パネル）

非復元リサンプリングによって 50 個の有意標本のサイズを 2～10 まで変化させ、それぞれの抽出誤差を計算することをさらに 500 回繰り返した結果が図 1 右パネルである。標本サイズの平方根に反比例して標本誤差が変化している。標本サイズに単純に回帰させた場合の決定係数は 0.662 である。

4 無作為抽出した場合の標本誤差

被引用回数が正規分布してれば、標準誤差は 7 である。有意抽出と同じサイズ、抽出回数で 500 回シミュレーションしてみると、標本誤差は最低で 5 最大で 8.9 、無作為標本の標本誤差の平均は 7 であり、有意抽出標本よりもずっと小さくなった。

5 補論

Y_{ij}は 期待値がμで、independently and identically distributed (iid) だと仮定する。またMとNは中心極限定理による十分な近似が得られるほど大きな値をとると仮定する。サンプルサイズが N_1, N_2 $(N_1 < N_2)$ のときの標本誤差を$\mathrm{s}_1(Y)$, $\mathrm{s}_2(Y)$とすると、証明したい命題は、

定理 2: $\mathrm{s}_1(Y) > \mathrm{s}_2(Y)$

である。サンプルサイズが N_1, N_2 のときの j 番目の標本平均を $\mathrm{m}_1(Y_j)$, $\mathrm{m}_2(Y_j)$ とすると、

$$\mathrm{s}_1(Y)^2 = \frac{\sum_{j=1}^{M}\left(\mathrm{m}_1(Y_j) - \mathrm{m}(Y)\right)^2}{M-1} = \frac{\sum_{j=1}^{M}\left(\mathrm{m}_1(Y_j)^2 - 2\mathrm{m}_1(Y_j)\cdot\mathrm{m}(Y) + \mathrm{m}(Y)^2\right)}{M-1}$$
$$= \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1(Y_j)^2 - \frac{2\cdot\mathrm{m}(Y)}{M-1}\sum_{j=1}^{M}\mathrm{m}_1(Y_j) + \frac{M}{M-1}\mathrm{m}(Y)^2 \qquad (6)$$

である。iid の仮定と中心極限定理より、$\sum_{j=1}^{M}\mathrm{m}_1(Y_j) = M\mu$であるから、これを式 (6) に代入して、

$$\mathrm{s}_1(Y)^2 = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1(Y_j)^2 - \frac{2M\cdot\mathrm{m}(Y)}{M-1}\mu + \frac{M}{M-1}\mathrm{m}(Y)^2 \qquad (7)$$

である。同様にして、

$$\mathrm{s}_2(Y)^2 \quad = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 - \frac{2M\cdot\mathrm{m}(Y)}{M-1}\mu + \frac{M}{M-1}\mathrm{m}(Y)^2 \qquad (8)$$

式(7)から式(8)をひくと、

$$\mathrm{s}_1(Y)^2 - \mathrm{s}_2(Y)^2 = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 - \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 \qquad (9)$$

である。

Y_{ij}の分散をvとすると、中心極限定理により、$\mathrm{m}_1(Y_j)$の分散は

$$\mathrm{v}(\mathrm{m}_1(Y_j)) = \frac{v}{N_1} = \frac{\sum_{j=1}^{M}(\,\mathrm{m}_1(Y_j)-\mu)^2}{M-1} = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 - \frac{M}{M-1}\mu^2 \qquad (10)$$

であり、同様に$\mathrm{m}_2(Y_j)$の分散は

$$\mathrm{v}(\mathrm{m}_2(Y_j)) = \frac{v}{N_2} = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 - \frac{M}{M-1}\mu^2 \qquad (11)$$

である。$N_1 < N_2$だから、式(10)、式(11)より、

$$\begin{aligned}
&\frac{v}{N_1} && > \frac{v}{N_2}\\
&\mathrm{v}(\mathrm{m}_1(Y_j)) && > \mathrm{v}(\mathrm{m}_2(Y_j))\\
&\frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 - \frac{M}{M-1}\mu^2 && > \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 - \frac{M}{M-1}\mu^2\\
&\frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 && > \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2\\
&\frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 - \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 && > 0 \qquad (12)
\end{aligned}$$

である。式(9)と式(12)より、

$$\begin{aligned}
&\mathrm{s}_1(Y)^2 - \mathrm{s}_2(Y)^2 = \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_1\,(Y_j)^2 - \frac{1}{M-1}\sum_{j=1}^{M}\mathrm{m}_2\,(Y_j)^2 && > 0\\
&\mathrm{s}_1(Y)^2 && > \mathrm{s}_2(Y)^2\\
&\mathrm{s}_1(Y) && > \mathrm{s}_2(Y) \qquad (13)
\end{aligned}$$

以上で証明終わり。