

社会調査 第5回・第6回

早稲田大学政治経済学部

2018年度秋学期

西郷浩(前半)・土屋隆裕(後半)

本日の目標

- 第5回講義
 - 層化抽出のシミュレーション
 - 層の決定
 - 層化の効果についての評価
 - 層化抽出の補足
- 第6回講義
 - 推定段階における補助情報の利用
 - 比推定量

第5回 層化抽出のシミュレーション

- 層の決定

- 原則

- 推定精度向上の観点

- 層内の散らばりは小さくする。
 - 層間の散らばり(総平均の散らばり)は大きくする。

- 分析(表章)の観点

- 分析(表章)に必要なグループは、標本の設計の段階で層としておく。
 - » 事後層化という方法もある。しかし、特定の(事後)層内の標本サイズが小さくなる可能性がある。

層化抽出のシミュレーション

－ 実際

- 補助変数の大きさを参考に、層への標本サイズの配分を決める。
- 最適配分 (Neyman 配分)

$$- n_h = n \times \frac{N_h S_h}{\sum_{h=1}^H N_h S_h} \propto N_h S_h$$

» 標本サイズ $n = \sum_{h=1}^H n_h$ を一定として母平均の推定量

$\hat{\mu} = \sum_{h=1}^H W_h \bar{y}_h$ の分散 $V(\hat{\mu}) = \sum_{h=1}^H W_h^2 \frac{S_h^2}{n_h} (1 - f_h)$ が最小になる。

- － 実際には、調査対象変数 y そのものではなく、補助変数で計算するので、「最適」とは限らない。

層化抽出のシミュレーション

- シミュレーション

- 層化抽出法1

- 層1: 作付面積250ha以上。 $f_1 = 4/4$.
 - 層2: 作付面積250ha未満。 $f_2 = 6/65$.

- 層化抽出法2

- 層1: 作付面積250ha以上。 $f_1 = 4/4$.
 - 層2: 作付面積100ha以上250ha未満。 $f_2 = 3/10$.
 - 層3: 作付面積100ha未満。 $f_3 = 3/55$.

層化抽出のシミュレーション

– 層化抽出法3 (Neyman配分)

- 層1: 作付面積250ha以上。 $f_1 = 2/4$.
- 層2: 作付面積100ha以上250ha未満。 $f_2 = 2/10$.
- 層3: 作付面積100ha未満。 $f_3 = 6/55$.

• 層化の効果についての評価

– 層化によって推定効率を格段に向上できる。

- デザイン効果 $Deff = \frac{V(\hat{\mu}_{other})}{V(\hat{\mu}_{SRS})}$
 - ただし、分子・分母で標本の大きさ n は一定とする。

層化抽出のシミュレーション

– シミュレーションの値から計算した $Deff$

- 表1: シミュレーションにおける $Deff$ の値

抽出方法	SRS	層化1	層化2	層化3
$Deff$	1.00	0.45	0.13	0.09

– シミュレーションで層化が効果的だった理由

- 母集団における収穫量 y の分布が大きく歪んでいる。
 - ヒストグラムの右裾に位置する抽出単位が標本に含まれるか否かで、(SRSにおける)標本平均の値が大きく変動する。
 - » 層化抽出による変動抑制の効果が大きい。
- 調査対象変数 y と相関の強い補助変数がある。

層化抽出の補足

- Between variation and Within variation

$$\begin{aligned} - (N - 1)S^2 &= \sum_{h=1}^H \sum_{i=1}^{N_h} (y_{hi} - \mu)^2 \\ &= \sum_{h=1}^H \sum_{i=1}^{N_h} \{(y_{hi} - \mu_h) + (\mu_h - \mu)\}^2 \\ &= \sum_{h=1}^H \left\{ N_h(\mu_h - \mu)^2 + \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2 \right\} \\ &= \sum_{h=1}^H \{ N_h(\mu_h - \mu)^2 + (N_h - 1)S_h^2 \} \\ &\equiv \textit{Between Variation} + \textit{Within Variation} \end{aligned}$$

層化抽出の補足

– Use the following approximation.

- $W_h \equiv \frac{N_h}{N} \approx \frac{N_h}{N-1} \approx \frac{N_h-1}{N-1}.$

– Then the population variance is written as

- $S^2 \approx \sum_{h=1}^H W_h (\mu_h - \mu)^2 + \sum_{h=1}^H W_h S_h^2$

– Thus the variance of the sample mean under SRS is approximated by

- $V(\hat{\mu}_{SRS}) = V(\bar{y}_{SRS})$
 $\approx \frac{1-f}{n} \left\{ \sum_{h=1}^H W_h (\mu_h - \mu)^2 + \sum_{h=1}^H W_h S_h^2 \right\}.$

層化抽出の補足

– The variance of the population estimator under stratified sampling is

- $V(\hat{\mu}_{st}) = \sum_{h=1}^H W_h^2 \frac{S_h^2}{n_h} (1 - f_h).$

– The design effect of $\hat{\mu}_{st}$ is given by

- $Def f \approx \frac{\sum_{h=1}^H W_h^2 \frac{S_h^2}{n_h} (1 - f_h)}{\frac{1-f}{n} \{ \sum_{h=1}^H W_h (\mu_h - \mu)^2 + \sum_{h=1}^H W_h S_h^2 \}}.$

層化抽出の補足

– With proportional allocation,

- $n_h = W_h n \propto N_h$.

- $Def f \approx \frac{\sum_{h=1}^H W_h S_h^2}{\sum_{h=1}^H W_h (\mu_h - \mu)^2 + \sum_{h=1}^H W_h S_h^2} = \frac{WV}{BV + WV} < 1$

- unless $\mu_h = \mu$ for all h .

第6回 比推定量・回帰推定量

- 推定段階における補助情報の利用
 - 比推定量
 - 回帰推定量
 - 差推定量
 - キャリブレーション
 - 事後層化

比推定量

- 比推定量の考え方

- 調査対象変数と比例関係の強い補助情報

- 例：収穫量と作付面積

- MS Excel ファイル(層化抽出法シミュレーション)の散布図を参照のこと。

- 比例関係の利用

- SRSのもとでの標本平均と母平均の関係

- $\frac{\bar{y}}{\mu_y} \approx \frac{\bar{x}}{\mu_x} \Leftrightarrow \mu_y \approx \frac{\bar{y}}{\bar{x}} \mu_x = \bar{y} \left(\frac{\mu_x}{\bar{x}} \right)$

- $\frac{\bar{y}}{\mu_y} \approx \frac{\bar{x}}{\mu_x} \Leftrightarrow \frac{N\bar{y}}{N\mu_y} \approx \frac{N\bar{x}}{N\mu_x} \Leftrightarrow \frac{N\bar{y}}{\tau_y} \approx \frac{N\bar{x}}{\tau_x} \Leftrightarrow \tau_y \approx \frac{\bar{y}}{\bar{x}} \tau_x$

比推定量

- 比推定量

- 母平均の推定量

- $\hat{\mu}_{ratio} = \frac{\bar{y}}{\bar{x}} \mu_x$

- 補助変数の母平均 μ_x は既知とする。

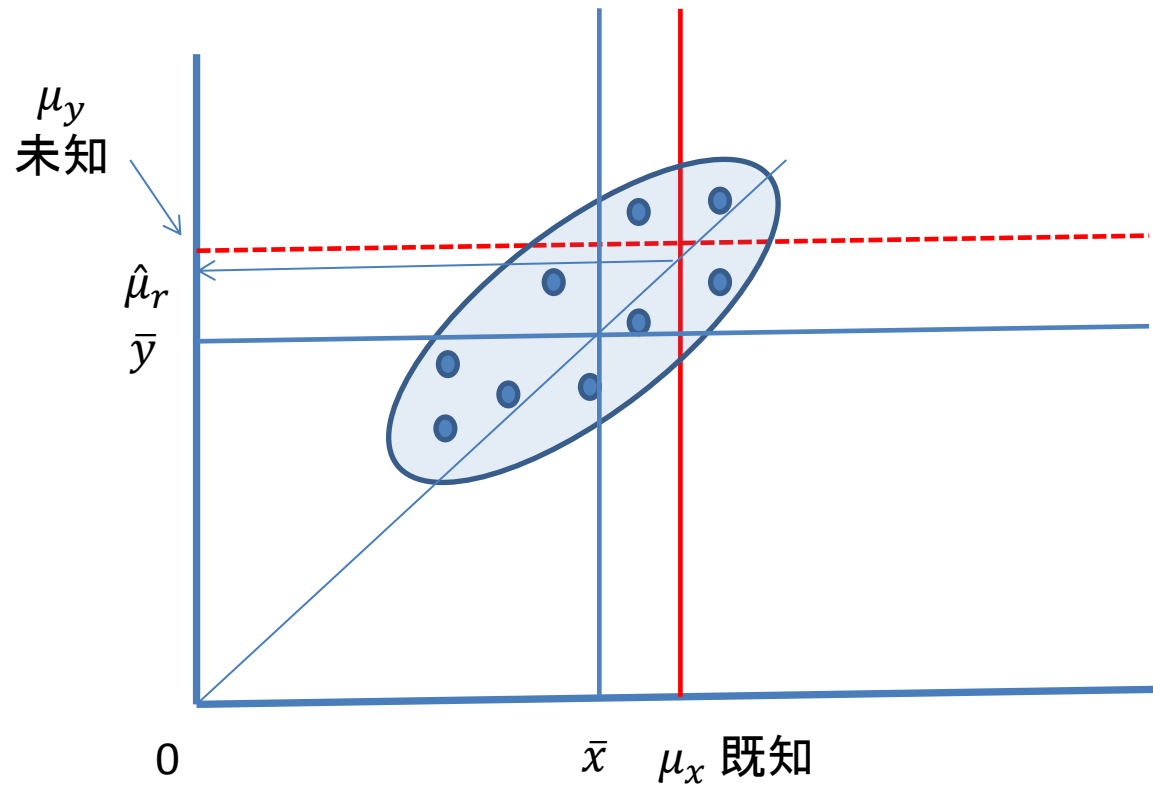
- 母集団合計の推定量

- $\hat{t}_{ratio} = \frac{\bar{y}}{\bar{x}} \tau_x = N \hat{\mu}_{ratio}$

- 補助変数の母集団合計 τ_x は既知とする。

比推定量

- 図式的な説明



比推定量

- 母平均の比推定量の分散とその推定量など
 - 分散

- $V(\hat{\mu}_{ratio}) \approx \frac{1-f}{n} \times \frac{1}{N-1} \sum_{i=1}^N (y_i - \rho x_i)^2 \equiv \frac{1-f}{n} S_{ratio}^2$
 - ただし、 $\rho = \frac{\mu_y}{\mu_x} = \frac{\tau_y}{\tau_x}$.

- 分散の推定量(標準誤差の二乗)

- $\hat{V}(\hat{\mu}_{ratio}) = \frac{1-f}{n} \hat{S}_{ratio}^2$
 - ただし、 $\hat{S}_{ratio}^2 \equiv \frac{1}{n-1} \sum_{i \in sample} (y_i - \hat{\rho} x_i)^2$ 、 $\hat{\rho} = \frac{\bar{y}}{\bar{x}}$

比推定量

－ 標準誤差 (分散の推定量の平方根)

$$\bullet \sqrt{\frac{1-f}{n} \hat{S}_{ratio}^2}$$

－ 母平均の信頼係数0.95の信頼区間

$$\bullet \hat{\mu}_{ratio} \pm 1.96 \sqrt{\frac{1-f}{n} \hat{S}_{ratio}^2}$$

• 母集団合計の信頼係数0.95の信頼区間

$$\bullet \hat{t}_{ratio} \pm 1.96N \sqrt{\frac{1-f}{n} \hat{S}_{ratio}^2}$$

比推定量

- シミュレーション
 - 母集団
 - そばを生産している長野県の市区町村
 - 推定対象
 - 市町村当たり平均そば収穫量(t)
 - 標本抽出
 - SRS
 - 推定方法
 - 比推定

比推定量

－ 比推定の効果

- 推定精度は層化抽出法3と同程度に高い。
 - － 補助変数と対象変数との比例関係が強いため。
- 捕捉率が低い。
 - － 標準誤差の推定値が過小推定になっている。
 - » 比推定量には偏りがある。
 - 標本の大きさ n が大きくなるにつれて、偏りが小さくなる。
 - 標本の大きさが小さいと、偏りが顕著になる。
 - » 標準誤差の計算において、偏りを無視している。
 - » 母集団における比 ρ が標本における比 $\hat{\rho}$ に置き換えられている

比推定量

- 捕捉

- A Taylor series expansion around (μ_y, μ_x) yields

- $$\hat{\mu}_{ratio} = \frac{\bar{y}}{\bar{x}} \mu_x \approx \mu_x \left(\frac{\mu_y}{\mu_x} + \left(\frac{1}{\mu_x} - \frac{\mu_y}{\mu_x^2} \right) (\bar{y} - \mu_y) \right)$$
$$= \mu_y + \{(\bar{y} - \mu_y) - \rho(\bar{x} - \mu_x)\} = \mu_y + (\bar{y} - \rho\bar{x}),$$
 - up to the first order term.

- The expectation of $\hat{\mu}_{ratio}$

- $E(\hat{\mu}_{ratio}) \approx 0.$
 - Note that $E(\bar{y}) = \mu_y, E(\bar{x}) = \mu_x, \rho = \mu_y/\mu_x.$

比推定量

– The variance of $\hat{\mu}_{ratio}$

- $V(\hat{\mu}_{ratio}) = E\{(\hat{\mu}_{ratio} - E(\hat{\mu}_{ratio}))^2\} \approx E\{(\bar{y} - \rho\bar{x})^2\}.$
- Since $E(\bar{y} - \rho\bar{x}) = 0$, $E\{(\bar{y} - \rho\bar{x})^2\} = V(\bar{y} - \rho\bar{x}).$
- Observe that

$$- \bar{y} - \rho\bar{x} = \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \rho x_i) \equiv \frac{1}{n} \sum_{i \in \text{sample}} e_i \equiv \bar{e}.$$

$$\gg \text{Note that } \mu_e = \frac{1}{N} \sum_{i=1}^N e_i = \frac{1}{N} \sum_{i=1}^N (y_i - \rho x_i) = 0.$$

- Finally, we have

$$- V(\hat{\mu}_{ratio}) \approx V(\bar{e}) = \frac{1-f}{n} S_{ratio}^2,$$

$$\gg \text{where } S_{ratio}^2 \equiv \frac{1}{N-1} \sum_{i=1}^N (e_i - \mu_e)^2 = \frac{1}{N-1} \sum_{i=1}^N e_i^2.$$

比推定量

– Variance estimation

- $\hat{V}(\hat{\mu}_{ratio}) = \frac{1-f}{n} \hat{S}_{ratio}^2$,
 - where $\hat{S}_{ratio}^2 = \frac{1}{n-1} \sum_{i \in sample} (y_i - \hat{\rho} x_i)^2$.
 - » If ρ is known, S_{ratio}^2 is estimated by
 - $\frac{1}{n-1} \sum_{i \in sample} (e_i - \bar{e})^2$,
 - where $e_i = y_i - \rho x_i$.
 - » Since it is unknown, we replace ρ with $\hat{\rho} = \bar{y}/\bar{x}$.
 - Note that $\sum_{i \in sample} (y_i - \hat{\rho} x_i) = 0$.