

社会調査 第3回・第4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3回授業
 - 単純無作為抽出法(続き)
 - シミュレーション
 - 理論
 - 系統抽出法
- 第4回授業
 - 母比率の推定
 - 層化抽出法
 - 定義・性質

第3回 単純無作為抽出法

- デザイン
 - 標本として可能な抽出単位(母集団の構成要素)の組み合わせを等確率で選ぶ抽出方法
- 実行方法(一様乱数の使用は前提とする。)
 - 複数あり
 - 例
 - 母集団の N 個の抽出単位のそれぞれに、区間 $(0,1)$ の一様乱数を発生する。大きい方 n 個に対応する抽出単位を標本とする。

単純無作為抽出法

- シミュレーション
 - 母集団
 - 抽出単位: 長野県内のそばを生産する69市町村
 - 変数: 抽出単位内のそばの収穫量
 - 推定対象:
 - 長野県内のそばの収穫量の一市町村当たりの平均値
 - 69倍すれば長野県内のそばの収穫量の合計になる。
 - サンプルサイズ: $n = 10$ (市町村)
 - 抽出方法: 単純無作為抽出法

単純無作為抽出法

- 推定

- 点推定:

- 母平均 $\mu \leftarrow$ 標本平均 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
 - 母集団の合計 $\tau \leftarrow$ 標本平均の N 倍 $\hat{\tau} = N\bar{y}$
 - 母分散 $S^2 \leftarrow$ 標本分散 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$
 - 標準誤差の平方 $\leftarrow \frac{s^2}{n} (1 - f)$
 - ただし、 $f = n/N$
 - 標準誤差 $\leftarrow \sqrt{\frac{s^2}{n} (1 - f)}$

単純無作為抽出法

－ 区間推定：

- 信頼係数0.95の母平均 μ の信頼区間

$$- \bar{y} \pm 1.96 \sqrt{\frac{s^2}{n} (1 - f)}$$

» 1.96 は、標準正規分布の上側0.025点

- － 区間推定値の構成は、「標本平均の標本分布が正規分布で近似できる」という性質に依拠する。

- 信頼係数0.95の母集団合計 τ の信頼区間

$$- N\bar{y} \pm 1.96N \sqrt{\frac{s^2}{n} (1 - f)}$$

單純無作為抽出法

- 補足(数理)

- Sample membership indicators

- $I_i = \begin{cases} 1 & i \in \text{sample} \\ 0 & i \notin \text{sample} \end{cases}$

- Note that $I_i^2 = I_i$ for all i .

- Under SRS with size n ,

- $\sum_{i=1}^N I_i = n$.

- This implies that $E\left(\sum_{i=1}^N I_i\right) = \sum_{i=1}^N E(I_i) = n$.

- By symmetry, $E(I_1) = E(I_2) = \cdots = E(I_N)$.

- Thus, $E(I_i) = n/N = f = \Pr(I_i = 1) = \Pr(i \in \text{sample})$.

- Observe that $V(I_i) = f(1 - f)$.

單純無作為抽出法

– Joint inclusion probabilities

$$\begin{aligned} \bullet \Pr(i, j \in \text{sample}) &= \Pr(I_i = 1, I_j = 1) \\ &= \Pr(I_i = 1) \Pr(I_j = 1 | I_i = 1) = \frac{n}{N} \times \frac{n-1}{N-1} \\ &= E(I_i I_j) \end{aligned}$$

– The covariance between I_i and I_j is

$$\begin{aligned} \gg \text{Cov}(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) = \frac{n}{N} \times \frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{f(1-f)}{N-1}. \end{aligned}$$

單純無作為抽出法

– The sample mean

- $\bar{y} = \frac{1}{n} \sum_{i \in \text{sample}} y_i = \frac{1}{n} \sum_{i=1}^N I_i y_i$

- Expectation:

- $E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N E(I_i) y_i = \frac{1}{n} \sum_{i=1}^N \left(\frac{n}{N}\right) y_i = \frac{1}{N} \sum_{i=1}^N y_i = \mu.$

- Variance:

- $$\begin{aligned} V(\bar{y}) &= \left(\frac{1}{n}\right)^2 \left\{ \sum_{i=1}^N y_i^2 V(I_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N y_i y_j \text{Cov}(I_i, I_j) \right\} \\ &= \left(\frac{1}{n}\right)^2 \left\{ f(1-f) \sum_{i=1}^N y_i^2 - 2 \frac{f(1-f)}{N-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^N y_i y_j \right\} \\ &= \frac{f(1-f)}{n^2(N-1)} \left\{ (N-1) \sum_{i=1}^N y_i^2 - 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N y_i y_j \right\} \end{aligned}$$

單純無作為抽出法

– Observe that

$$\begin{aligned} & \gg (N-1) \sum_{i=1}^N y_i^2 - 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N y_i y_j \\ &= (N-1) \sum_{i=1}^N y_i^2 - \sum_{i=1}^N \sum_{j \neq i}^N y_i y_j \\ &= N \sum_{i=1}^N y_i^2 - \sum_{i=1}^N \sum_{j=1}^N y_i y_j \\ &= N \sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N y_i \right)^2 \\ &= N \left\{ \sum_{i=1}^N y_i^2 - N \mu^2 \right\} \\ &= N \sum_{i=1}^N (y_i - \mu)^2 = N(N-1)S^2 \end{aligned}$$

單純無作為抽出法

– Finally, we obtain

$$\gg V(\bar{y}) = \frac{s^2}{n}(1 - f).$$

– Variance estimator

- The sample variance $s^2 = \frac{1}{n-1} \sum_{i \in \text{sample}} (y_i - \bar{y})^2$.
- Below we show that s^2 is unbiased for S^2 .

– First, rewrite s^2 as

$$\begin{aligned} \gg s^2 &= \frac{1}{n-1} \sum_{i \in \text{sample}} \{(y_i - \mu) - (\bar{y} - \mu)\}^2 \\ &= \frac{1}{n-1} \left\{ \sum_{i \in \text{sample}} (y_i - \mu)^2 - n(\bar{y} - \mu)^2 \right\} \\ &= \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \mu)^2 - (\bar{y} - \mu)^2 \right\} \end{aligned}$$

單純無作為抽出法

- The first term in the curly bracket is the sample mean of $(y_i - \mu)^2$. The expectation under SRS is then given by

$$\gg E \left\{ \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \mu)^2 \right\} = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 = \frac{N-1}{N} S^2.$$

- The expectation of the second term in it is the variance of the sample mean.

$$\gg E\{(\bar{y} - \mu)^2\} = V(\bar{y}) = \frac{S^2}{n} (1 - f) = \left(\frac{1}{n} - \frac{1}{N} \right) S^2.$$

- The expectation of s^2 is then obtained by

$$\gg E(s^2) = \frac{n}{n-1} \left(1 - \frac{1}{N} - \frac{1}{n} + \frac{1}{N} \right) S^2 = S^2.$$

単純無作為抽出法

- 補足（復元抽出の場合との比較）
 - Under SRS with replacement,
 - $E(\bar{y}) = \mu$
 - $V(\bar{y}) = \sigma^2/n$, where $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$
 - $E(s^2) = \sigma^2$
 - Level-0.95 confidence interval of μ : $\bar{y} \pm 1.96 \sqrt{\frac{s^2}{n}}$
 - When $f \approx 0$, SRS without replacement is essentially equivalent to SRS with replacement.

系統抽出法

- 抽出方法

- 例：大きさ 1000 の母集団から、大きさ20の標本を抽出する。

- 手順

- 抽出間隔を計算する。 $N/n = 1000/20 = 50$.
 - 1から50までの整数型一様乱数を1つ発生する。例：31
 - 番号31、 $31+50$ 、 $31+2*50$ 、...、 $31+(20-1)*50$ の抽出単位を標本とする。

- 別名：等間隔抽出法

系統抽出法

- 性質
 - 母平均・母集団合計を偏りなく推定できる。
 - 推定精度が母集団の並び方に依存する。
 - ランダムに並んだ母集団
 - SRSと同程度の推定精度
 - 小さい順に並んだ母集団
 - SRSよりも高い推定精度
 - 並び方に周期性のある母集団
 - SRSよりも低い推定精度

系統抽出法

- SRSと比較した場合の長短
 - 長所
 - 乱数の発生が1つですむ。
 - 母集団の並び方がランダムであればSRSの簡便法として利用できる。
 - 確率比例抽出法への拡張が容易である。
 - 短所
 - 標準誤差の推定が原理的に不可能である。
 - 母集団の並び方がランダムであれば、SRSと同じ方法で標準誤差を推定できる。

系統抽出法

- 実際
 - 様々な公的統計で利用されている。

第4回 母比率の推定

- 母比率の推定
 - 変数 y_i が質的変数(属性変数)の場合
 - 例
 - 千代田区内に勤務している人たちの住所
- » $y_i = \begin{cases} \text{新宿区} \\ \text{練馬区} \\ \vdots \end{cases}$
- そのままでは加減乗除ができない。

母比率の推定

- 0-1変数の導入

- もとの変数 y_i に付随して、人工的な変数を定義する。

- $y'_i = \begin{cases} 1 & y_i = a \text{ のとき} \\ 0 & y_i \neq a \text{ のとき} \end{cases}$

- ただし、 a は y_i の取りうる値のうちの特定の一つ。

- 例:

- 千代田区内に勤務している人たちのうち、新宿区に居住する人

- » $y'_i = \begin{cases} 1 & y_i = \text{新宿 のとき} \\ 0 & y_i \neq \text{新宿 のとき} \end{cases}$

母比率の推定

- 0-1変数の母平均と母集団合計

- 母平均

- $\mu_{y'} = \frac{1}{N} \sum_{i=1}^N y'_i = \frac{y_i=a \text{ となる個体の数}}{N} \equiv \pi_a$

- これまでの議論と形式的に同型である。

- » y'_i で母平均を推定すれば、母比率 π_a を推定できる。

- 後述の通り、公式はいつそう簡便になる。

- 母集団合計

- $\tau_{y'} = \sum_{i=1}^N y'_i = y_i = a \text{ となる個体の数} \equiv N_a$

母比率の推定

- 0-1変数で計算した母分散・標本分散・標準誤差

$$- S_{y'}^2 = \frac{N}{N-1} \pi_a (1 - \pi_a)$$

- 計算

$$\begin{aligned} - S_{y'}^2 &= \frac{1}{N-1} \sum_{i=1}^N (y'_i - \pi_a)^2 = \frac{1}{N-1} \{ \sum_{i=1}^N y_i'^2 - N\pi_a^2 \} \\ &= \frac{1}{N-1} \{ \sum_{i=1}^N y'_i - N\pi_a^2 \} = \frac{1}{N-1} \{ N\pi_a - N\pi_a^2 \} = \frac{N}{N-1} \pi_a (1 - \pi_a). \end{aligned}$$

$$- s_{y'}^2 = \frac{n}{n-1} \hat{\pi}_a (1 - \hat{\pi}_a)$$

- ただし、 $\hat{\pi}_a = \bar{y}' = \frac{1}{n} \sum_{i \in \text{sample}} y'_i =$ 標本における a の比率

$$- \sqrt{\frac{s_{y'}^2}{n}} (1 - f) = \sqrt{\frac{\hat{\pi}_a (1 - \hat{\pi}_a)}{n-1}} (1 - f)$$

母比率の推定

- 信頼係数0.95の信頼区間
 - 属性 a を持つ個体の母比率

- $\hat{\pi}_a \pm 1.96 \sqrt{\frac{\hat{\pi}_a(1-\hat{\pi}_a)}{n-1}} (1-f)$

- 属性 a を持つ個体の総数

- $N\hat{\pi}_a \pm 1.96N \sqrt{\frac{\hat{\pi}_a(1-\hat{\pi}_a)}{n-1}} (1-f)$

母比率の推定

- シミュレーション
 - 母集団の大きさ: 100
 - 変数: 居住地
 - 架空のデータ
 - 推定対象: 都内(23区)居住者の比率・人数
 - 標本の大きさ: 10

層化抽出法

- 補助情報

- － 標本抽出に先立って、母集団の抽出単位に関して利用可能な情報

- 例：

- － 過去の作付面積

- － 経済センサス実施年の業種・従業者規模

- 利用方法

- － 標本抽出段階：層化抽出、確率比例抽出

- － 推定段階：比推定、回帰推定

層化抽出法

- 層化の動機

- 単純無作為抽出法

- 比較的大きな値を持つ抽出単位がかなり多く抽出されるという可能性がある。
 - 特定の標本における変数の分布が、母集団におけるそれからかけ離れていることがある。
 - 「変数の分布が、母集団におけるそれからかけ離れている」標本を、なるべく少なくする方法は？
 - すなわち、「変数の分布が母集団のそれに近いという意味でバランスの良い」標本を得るための工夫は？

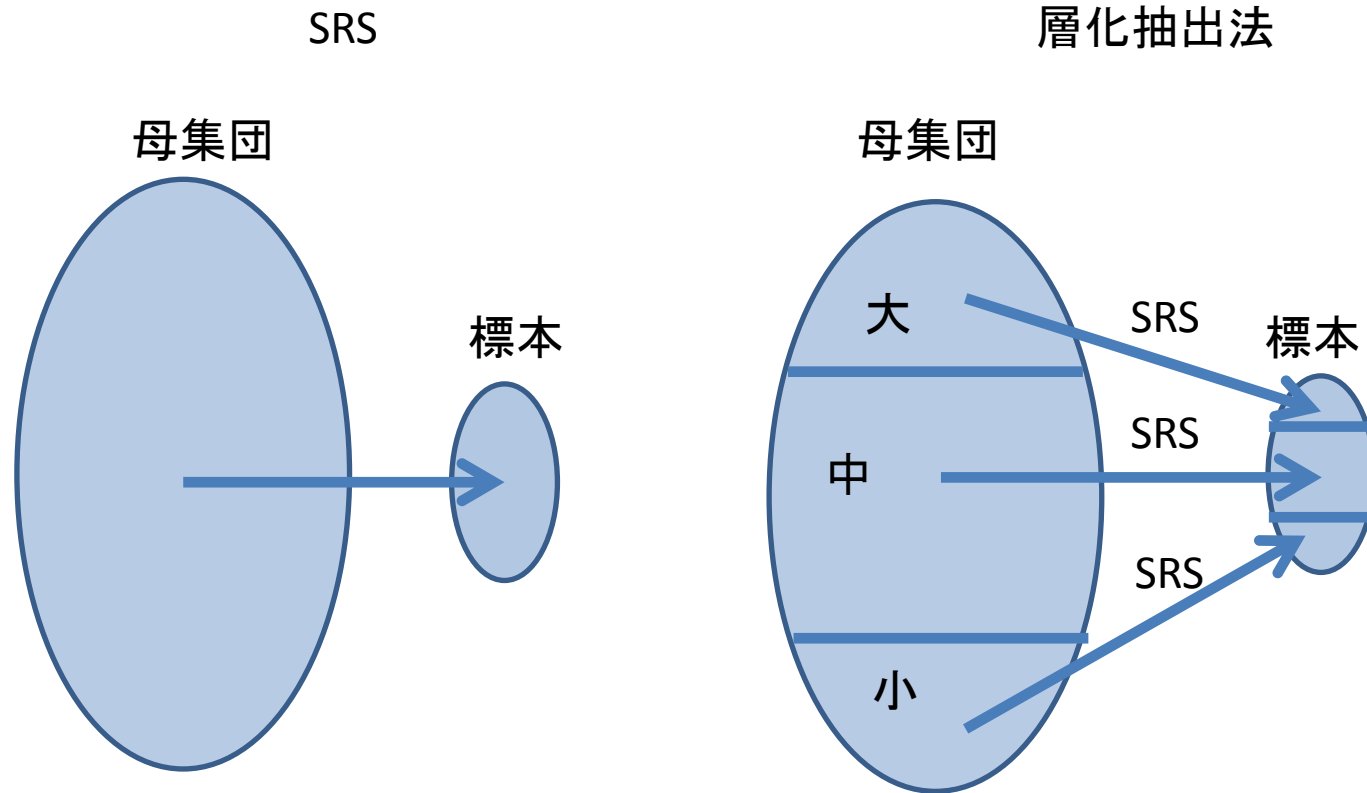
層化抽出法

－ 層化（層別）抽出法

- （理想形を示すための非現実的な仮定）
 - － 母集団における変数の値 y_i がすべて既知である。
- 変数の値 y_i の大・中・小、母集団を3つの部分母集団に分割する。
 - － 部分母集団を「層」と呼ぶ。
 - － 分割とは、(1)部分母集団に重複がない、(2)部分母集団を合併すれば母集団になる、ということの意味する。
- それぞれの層から単純無作為抽出する。
- それらを合わせて1つの標本とする。
 - － 大・中・小の抽出単位が決められた数で必ず抽出される。

層化抽出法

- SRSと層化抽出法のイメージ



層化抽出法

- 層化抽出法

- (現実的な仮定)

- これから調査する変数の値 y_i は未知である。
 - これまでの調査や行政記録から、抽出単位 i についての補助情報がある。

- 例:

- » 事業所調査:

- 所在地、業種、従業者規模、経済センサス実施時の売上高、など
- これから調査する変数の値 y_i の大小になるべく対応するように、層を構成する。

層化抽出法

- 記号

- 層 h に属する抽出単位 i の変数の値: y_{hi}

- $h = 1, 2, \dots, H; (h)i = 1, 2, \dots, N_h.$

- 母集団合計と層合計との関係

- $\tau = \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^H (\sum_{i=1}^{N_h} y_{hi}) = \sum_{h=1}^H \tau_h$

- 母平均と層平均の関係

- $\mu = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi} = \frac{1}{N} \sum_{h=1}^H N_h \mu_h = \sum_{h=1}^H W_h \mu_h$

- $\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi} = \tau_h / N_h \quad (h = 1, 2, \dots, H).$

- $W_h = N_h / N$

層化抽出法

- 推定量

- 母集団合計

- $\hat{t} = \sum_{h=1}^H \hat{t}_h$

- ただし、 $\hat{t}_h = N_h \bar{y}_h$, $\bar{y}_h = \frac{1}{n_h} \sum_{i \in \text{sample in stratum } h} y_{hi}$

- 母平均

- $\hat{\mu} = \sum_{h=1}^H W_h \hat{\mu}_h$

- ただし、 $W_h = N_h/N$, $\hat{\mu}_h = \bar{y}_h$.

層化抽出法

- 推定量の期待値・分散、標準誤差の推定量
– 母集団合計

- $$\begin{aligned} E(\hat{\tau}) &= E\left\{\sum_{h=1}^H \hat{\tau}_h\right\} = \sum_{h=1}^H E(\hat{\tau}_h) = \sum_{h=1}^H E(N_h \hat{\mu}_h) \\ &= \sum_{h=1}^H N_h E(\hat{\mu}_h) = \sum_{h=1}^H N_h \mu_h = \sum_{h=1}^H \tau_h = \tau. \end{aligned}$$

- $$\begin{aligned} V(\hat{\tau}) &= V\left\{\sum_{h=1}^H \hat{\tau}_h\right\} = \sum_{h=1}^H V(\hat{\tau}_h) = \sum_{h=1}^H V(N_h \hat{\mu}_h) \\ &= \sum_{h=1}^H N_h^2 V(\hat{\mu}_h) = \sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} (1 - f_h). \end{aligned}$$

- $$- \text{ただし、} S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2, f_h = n_h / N_h.$$

- $$\sqrt{\sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} (1 - f_h)}$$

- $$- \text{ただし、} s_h^2 = \frac{1}{n_h - 1} \sum_{i \in \text{sample in stratum } h} (y_{hi} - \bar{y}_h)^2.$$

層化抽出法

— 母平均

- $E(\hat{\mu}) = E\{\sum_{h=1}^H W_h \hat{\mu}_h\} = \sum_{h=1}^H W_h E(\hat{\mu}_h) = \sum_{h=1}^H W_h E(\bar{y}_h) = \sum_{h=1}^H W_h \mu_h = \mu.$
- $V(\hat{\mu}) = V\{\sum_{h=1}^H W_h \hat{\mu}_h\} = \sum_{h=1}^H W_h^2 V(\hat{\mu}_h) = \sum_{h=1}^H W_h^2 \frac{s_h^2}{n_h} (1 - f_h).$
- $\sqrt{\sum_{h=1}^H W_h^2 \frac{s_h^2}{n_h} (1 - f_h)}.$