

社会調査 第7回・第8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7回講義
 - 回帰推定量
 - 差推定量
- 第8回講義
 - 集落抽出法
 - 多段抽出法
 - その他の話題

第7回 回帰推定量

- 推定段階における回帰式の利用

- 比推定

- 補助情報 x と調査対象変数 y との比例関係の利用

- 原点を通る直線関係を想定

- » $y \approx \rho x$

- 回帰推定

- 補助情報 x と調査対象変数 y との相関関係の利用

- 必ずしも原点を通るとは限らない直線関係を想定

- » $y \approx \alpha + \beta x$

回帰推定量

- 単純無作為抽出法のもとでの回帰推定量

$$- \hat{\mu}_{reg} = \hat{\alpha} + \hat{\beta}\mu_x = \bar{y} + \hat{\beta}(\mu_x - \bar{x})$$

- ただし、

$$- \hat{\beta} = \frac{\sum_{i \in \text{sample}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \text{sample}} (x_i - \bar{x})^2}$$

$$- \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

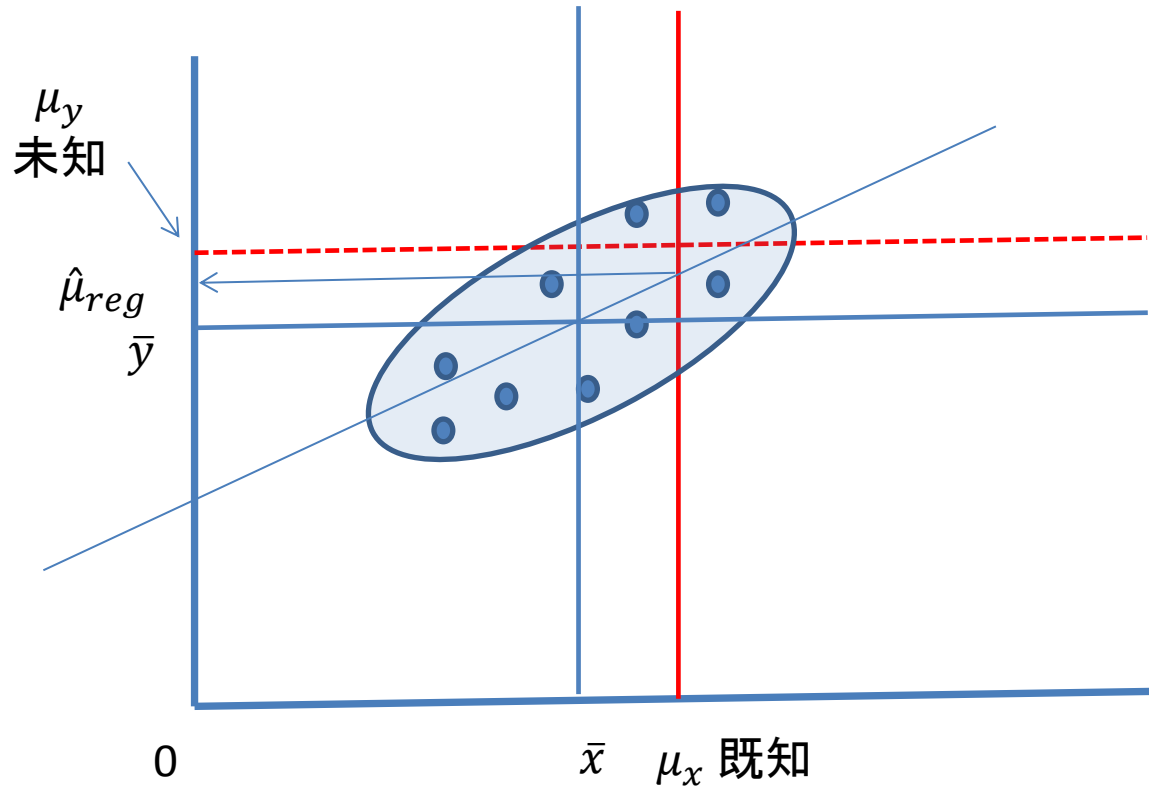
$$- \bar{y} = \frac{1}{n} \sum_{i \in \text{sample}} y_i, \bar{x} = \frac{1}{n} \sum_{i \in \text{sample}} x_i$$

$$- \mu_x = \frac{1}{N} \sum_{i=1}^N x_i$$

- 最小二乗法による(通常の)回帰式。

回帰推定量

- 図式的な説明



回帰推定量

- 回帰推定量の性質

- (近似的に) 偏りのない推定

- $\hat{\beta} \approx \beta = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$ が成り立つとする。

- $\hat{\mu}_{reg} = \bar{y} + \hat{\beta}(\mu_x - \bar{x}) \approx \bar{y} + \beta(\mu_x - \bar{x})$

- $E(\hat{\mu}_{reg}) \approx E\{\bar{y} + \beta(\mu_x - \bar{x})\} = E(\bar{y}) + \beta E(\mu_x - \bar{x})$
 $= \mu + 0 = \mu.$

- 回帰推定量の分散

- $V(\hat{\mu}_{reg}) = \frac{1-f}{n} S_e^2$

- ただし、 $S_e^2 = \frac{1}{N-1} \sum_{i=1}^N \{(y_i - \mu) - \beta(x_i - \mu_x)\}^2$

回帰推定量

– 推定量の分散の推定量

- $\hat{V}(\hat{\mu}_{reg}) = \frac{1-f}{n} s_e^2$

- ただし、 $s_e^2 = \frac{1}{n-1} \sum_{i \in sample} \{(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\}^2$

- » 回帰モデルからの類推で

- $s_e^2 = \frac{1}{n-2} \sum_{i \in sample} \{(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\}^2$ とする教科書もある。

– 信頼係数0.95の母平均信頼区間

- $\hat{\mu}_{reg} \pm 1.96 \sqrt{\frac{1-f}{n} s_e^2}$

回帰推定量

- シミュレーション
 - データ等：これまでと同じ。
 - 結果
 - 比推定値とほぼ同じ。
 - 推定精度 ($Deff = 0.08$ 、捕捉率 = 0.65)
 - » 母集団において、ほぼ、比例関係(相関関係の一種)が成り立っているため。
 - » 標準誤差の推定(したがって、信頼区間の構成)には、より大きなサンプルサイズが必要となる。

差(分)推定量

- 調査対象変数 y と補助情報 x との関係
 - 一般の関数
 - $y \approx g(x)$
 - 例
 - » $y \approx \alpha x^\beta$
- この近似関係を母平均の推定に活かす。
 - 残差(誤差)
 - $e_i = y_i - g(x_i)$ for all $i = 1, 2, \dots, N$.

差推定量

- 調査対象変数 y の母平均と $g(x)$ との関係。
 - $\mu = \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N \{g(x_i) + e_i\}$
 $= \frac{1}{N} \sum_{i=1}^N g(x_i) + \frac{1}{N} \sum_{i=1}^N e_i.$
 - 簡単のため、関数 $g(\cdot)$ が既知とする。
 - 補助情報 x_i ($i = 1, 2, \dots, N$) も既知とする。
 - しかし、誤差 $e_i = y_i - g(x_i)$ は標本でしか知ることができない。
 - つまり、未知。

差推定量

- 差推定量

- $\hat{\mu}_d = \frac{1}{N} \sum_{i=1}^N g(x_i) + \frac{1}{n} \sum_{i \in \text{sample}} e_i$

- 差推定量の性質

- (任意の関数 $g(x_i)$ について) 不偏推定が可能。

- $$\begin{aligned} E(\hat{\mu}_d) &= E \left\{ \frac{1}{N} \sum_{i=1}^N g(x_i) + \frac{1}{n} \sum_{i \in \text{sample}} e_i \right\} \\ &= \frac{1}{N} \sum_{i=1}^N g(x_i) + E \left\{ \frac{1}{n} \sum_{i \in \text{sample}} e_i \right\} \\ &= \frac{1}{N} \sum_{i=1}^N g(x_i) + \frac{1}{N} \sum_{i=1}^N e_i = \frac{1}{N} \sum_{i=1}^N \{g(x_i) + e_i\} = \mu. \end{aligned}$$

差推定量

– 差推定量の分散

$$\begin{aligned} \bullet V(\hat{\mu}_d) &= V\left\{\frac{1}{N}\sum_{i=1}^N g(x_i) + \frac{1}{n}\sum_{i \in \text{sample}} e_i\right\} \\ &= V\left\{\frac{1}{n}\sum_{i \in \text{sample}} e_i\right\} = \frac{1-f}{n} S_e^2, \end{aligned}$$

$$- \text{ただし、} S_e^2 = \frac{1}{N-1} \sum_{i=1}^N \left\{e_i - \frac{1}{N} \sum_{j=1}^N e_j\right\}^2$$

– 差推定量の分散の推定量

$$\bullet \hat{V}(\hat{\mu}_d) = \frac{1-f}{n} s_e^2$$

$$- \text{ただし、} s_e^2 = \frac{1}{n-1} \sum_{i \in \text{sample}} \left\{e_i - \frac{1}{n} \sum_{j \in \text{sample}} e_j\right\}^2$$

差推定量

– 信頼係数0.95の母平均の信頼区間

$$\bullet \hat{\mu}_d \pm 1.96 \sqrt{\frac{1-f}{n} s_e^2}$$

• 実際には、関数 $g(\cdot)$ が未知である。

– 標本から推定する。推定された関数: $\hat{g}(\cdot)$

$$\bullet \hat{\mu}_d = \frac{1}{N} \sum_{i=1}^N \hat{g}(x_i) + \frac{1}{n} \sum_{i \in \text{sample}} \hat{e}_i$$

– ただし、 $\hat{e}_i = y_i - \hat{g}(x_i)$

• 補助情報 x_i が、母集団に属するすべての抽出単位 $i = 1, 2, \dots, N$ について既知であることが前提である。

• 標準誤差が過小推定される。

差推定量

- シミュレーション
 - データ:これまでと同じ。
 - 調査対象変数 y と補助情報 x との関係の想定
 - $y = \alpha x^\beta \Leftrightarrow \log_{10} y = \log_{10} \alpha + \beta \log_{10} x$
 - 結果
 - 比推定量、回帰推定量とほぼ同じ結果
 - 推定精度 ($\text{Deff} = 0.11$ 、捕捉率 = 0.65)

差推定量

- 差推定量の特殊形

- 比推定量

- 想定されている関数形

- $y \approx \rho x$

- このとき、推定された残差の平均が0になる。

- $-\frac{1}{n} \sum_{i \in \text{sample}} \hat{e}_i = \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \hat{\rho} x_i) = \bar{y} - \hat{\rho} \bar{x} = 0.$

- したがって、差推定量が比推定量に等しくなる。

- $\hat{\mu}_d = \frac{1}{N} \sum_{i=1}^N \hat{\rho} x_i + \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \hat{\rho} x_i) = \hat{\rho} \frac{1}{N} \sum_{i=1}^N x_i = \hat{\mu}_r.$

差推定量

－ 回帰推定量

- 想定されている関数形

$$- y \approx \alpha + \beta x$$

- このとき、最小二乗法で推定された残差の平均が0になる。

$$- \frac{1}{n} \sum_{i \in \text{sample}} \hat{e}_i = \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0.$$

- したがって、差推定量が比推定量に等しくなる。

$$\begin{aligned} - \hat{\mu}_d &= \frac{1}{N} \sum_{i=1}^N \{\hat{\alpha} + \hat{\beta} x_i\} + \frac{1}{n} \sum_{i \in \text{sample}} (y_i - \hat{\rho} x_i) \\ &= \hat{\alpha} + \hat{\beta} \frac{1}{N} \sum_{i=1}^N x_i = \hat{\mu}_{reg}. \end{aligned}$$

差推定量

- 母集団の合計 $\tau = N\mu$ の推定
 - 母平均の推定量に母集団の大きさ N を乗じる。
 - 母集団合計の推定量
 - $\hat{\tau}_d = N\hat{\mu}_d$
 - 母集団合計の推定量の分散とその推定量
 - $V(\hat{\tau}_d) = N^2V(\hat{\mu}_d), \hat{V}(\hat{\tau}_d) = N^2\hat{V}(\hat{\mu}_d).$
 - 信頼係数0.95の信頼区間
 - $N\hat{\mu}_d \pm 1.96N\sqrt{\frac{1-f}{n}}s_e^2$

第8回 集落抽出法

- 抽出単位と統計単位

- 抽出単位

- 母集団から標本を抽出するときの、最小限の構成要素

- 例:

- » 大学の学籍簿を利用して、在籍者の標本を抽出する。
 - 母集団: その大学の全在籍者
 - 抽出単位: その大学の個々の在籍者
 - » 事業所母集団DBを利用して、事業所の標本を抽出する。
 - 母集団: わが国の全事業所
 - 抽出単位: わが国で操業する事業所

集落抽出法

－ 統計単位（集計単位）

- 調査結果を集計するときの集団の構成要素

- － 例：

- » 従業者規模別事業所数（従業者数の度数分布）

- 統計単位：事業所

- » 有権者の支持政党の分布

- 統計単位：有権者

－ 抽出単位と統計単位が一致しない場合がある。

- 例

- － 世帯を抽出単位として標本を抽出して、有権者を統計単位として支持政党の分布を調べる。

集落抽出法

- 集落抽出法

- 複数の統計単位の集合(集落 クラスター)を抽出単位として標本抽出する方法

- 例:

- 全国の高校3年生の平均的な学力を調べたい。
 - 高校(抽出単位)を無作為抽出する。
 - 抽出された高校の3年生全員に学力テストを実施する。
 - 学力テスト受けた生徒(統計単位)の試験結果を、(生徒一人当たりの)平均点や分散、度数分布表にまとめる。

集落抽出法

- 母集団の合計の推定
 - 記号の整理
 - 調査対象変数: y_{ai}
 - ただし、 y_{aj} は第 a 集落の第 i 統計単位の値をあらわす。
 - 集落(抽出単位)の総数: N
 - 集落 a の大きさ(集落 a に含まれる統計単位の総数): M_a
 - 簡単のため、すべての集落の大きさが等しいと仮定することもある。 $M_a = M$ for all a .
 - 集落 a の中の調査対象変数 y_{ai} の合計: $\tau_a = \sum_{i=1}^{M_a} y_{ai}$

集落抽出法

- 母集団の合計: $\tau = \sum_{a=1}^N \tau_a = \sum_{a=1}^N \sum_{i=1}^{M_a} y_{ai}$

– 標本抽出

- (集落の)単純無作為抽出法
 - 標本 sample は n 個の集落から成る。

– 推定の方針

- 集落当たりの母平均を $\mu = \frac{1}{N} \sum_{a=1}^N \tau_a$ とする。
 - 記号が異なるだけで、(集落当たりの)母平均の推定を N 倍すれば、母集団合計を推定できる。

集落抽出法

– 推定方法

- (集落当たりの)母平均の推定

$$- \bar{\tau} = \frac{1}{n} \sum_{a \in \text{sample}} \tau_a = \frac{1}{n} \sum_{a \in \text{sample}} \sum_{i=1}^{M_a} y_{ai}$$

- 推定量の分散

$$- V(\bar{\tau}) = \frac{1-f}{n} S_{\tau_a}^2$$

$$\gg \text{ただし、} S_{\tau_a}^2 = \frac{1}{N-1} \sum_{a=1}^N (\tau_a - \mu)^2$$

- 推定量の分散の推定量

$$- \hat{V}(\bar{\tau}) = \frac{1-f}{n} s_{\tau_a}^2$$

$$\gg \text{ただし、} s_{\tau_a}^2 = \frac{1}{n-1} \sum_{a \in \text{sample}} (\tau_a - \hat{\mu})^2$$

集落抽出法

- 統計単位当たりの母平均の推定

- 統計単位当たりの母平均

- $$\mu_{unit} = \frac{\sum_{a=1}^N \sum_{i=1}^{M_a} y_{ai}}{\sum_{a=1}^N M_a} = \frac{\sum_{a=1}^N \tau_a}{\sum_{a=1}^N M_a} = \frac{\frac{1}{N} \sum_{a=1}^N \tau_a}{\frac{1}{N} \sum_{a=1}^N M_a} \equiv \frac{\mu}{\mu_M}$$

- 統計単位当たりの母平均の推定量

- $$\hat{\mu}_{unit} = \frac{\frac{1}{n} \sum_{a \in sample} \tau_a}{\frac{1}{n} \sum_{a \in sample} M_a} = \frac{\bar{\tau}}{\bar{M}}$$

- 比推定で利用した方法を援用できる。

集落抽出法

– Taylor展開の利用

- $\hat{\mu}_{unit} \approx \mu_{unit} + \frac{1}{\mu_M} \times \frac{1}{n} \sum_{a \in sample} e_a$
 - ただし、 $e_a = \tau_a - \mu_{unit} M_a$.

– $\hat{\mu}_{unit}$ の期待値

- $E(\hat{\mu}_{unit}) \approx \mu_{unit}$
 - 近似的に偏りのない推定

– $\hat{\mu}_{unit}$ の分散

- $V(\hat{\mu}_{unit}) \approx \frac{1-f}{n\mu_M^2} S_e^2$
 - ただし、 $S_e^2 = \frac{1}{N-1} \sum_{a=1}^N e_a^2$

集落抽出法

– $\hat{\mu}_{unit}$ の分散の推定量

- $\hat{V}(\hat{\mu}_{unit}) \approx \frac{1-f}{n\mu_M^2} s_e^2$

- ただし、 $s_e^2 = \frac{1}{n-1} \sum_{a \in sample} \left\{ e_a - \frac{1}{n} \sum_{b \in sample} e_b \right\}^2$

– 信頼係数0.95の μ_{unit} の信頼区間

- $\hat{\mu}_{unit} \pm 1.96 \sqrt{\frac{1-f}{n\mu_M^2} s_e^2}$

– 未知母数を標本推定量で置き換える。

- $e_a = \tau_a - \mu_{unit} M_a \Rightarrow \hat{e}_a = \tau_a - \hat{\mu}_{unit} M_a$

- もし、 μ_M が未知なら、 $\mu_M \Rightarrow \bar{M}$

集落抽出法

- 例

- 米 Baltimore の約15000世帯から30世帯を無策抽出して、世帯員の数、性別、昨年診療を受けたかどうかを尋ねた。
- 以下の2つの比率を推定する。
 - 女の世界員の割合
 - 昨年診療を受けた世帯員の割合
- 出典 : Cochran, 1977, Sampling Techniques, 3rd ed., Wiley, pp. 66-68.

集落抽出法

－ 結果

- 女の世帯員の割合

- － 推定値:0.49

- － 推定量の分散の推定量:0.00114

- » 集落抽出であることを無視した場合:0.00240

- 婚姻によって世帯が形成されることが多かったので、多くの世帯において、男女とも少なくともひとりずついることが多い。その結果、世帯(集落)間の女の数の散らばりが小さい。

- それを無視して分散を推定すると過大推定になる。

集落抽出法

- 昨年診療を受けた世帯員の割合
 - 推定値:0.28
 - 推定量の分散の推定量:0.00521
 - » 集落抽出であることを無視した場合:0.00197
- 世帯員のひとりでも診療を受ければ、他の世帯員も診療を受ける可能性が高くなる(世帯内感染、あるいは世帯員の受診に関する考え方の類似)。その結果、世帯(集落)間での受診者数の散らばりが大きくなる。
 - それを無視して分散を推定すると過小推定になる。

集落抽出法

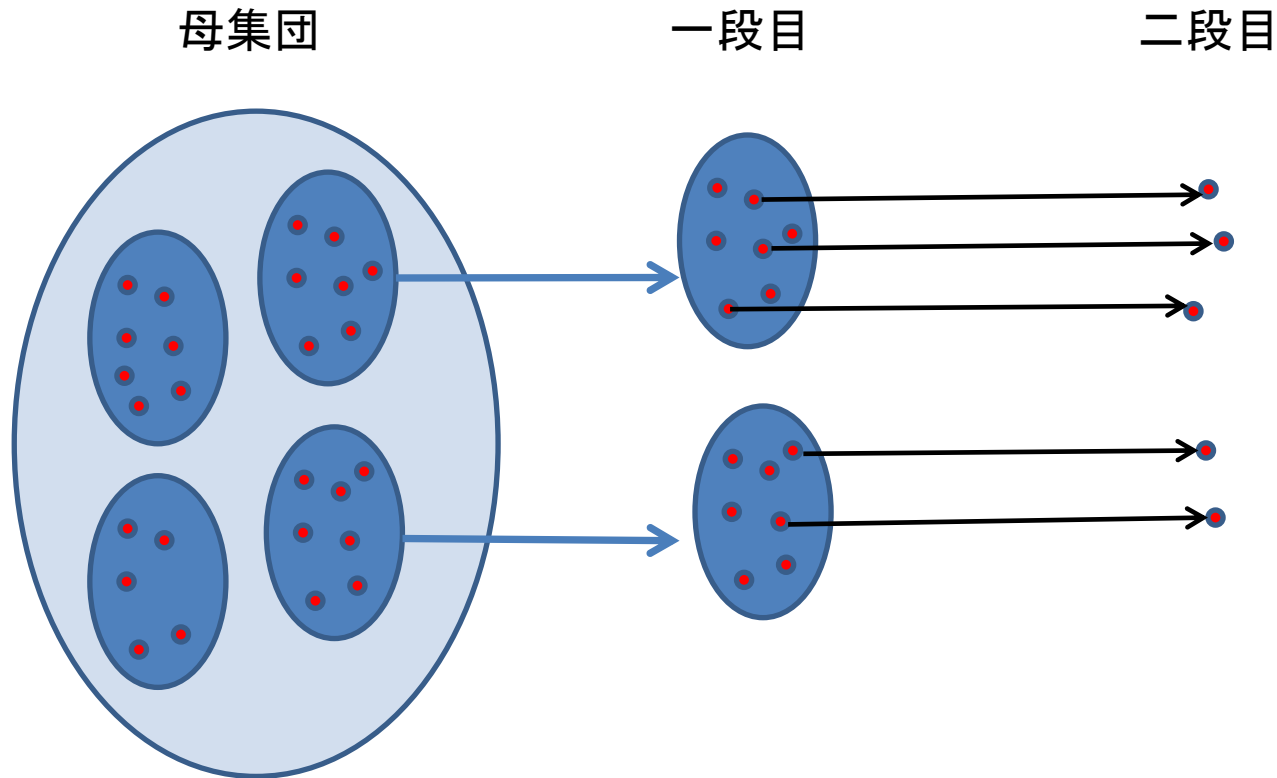
- 集落抽出法 vs 単純無作為抽出法
 - (統計単位という意味での) 標本の大きさが同じなら:
 - 集落抽出法のもとでの母平均の推定量の分散
＞ SRSのもとでの母平均の推定量の分散
 - 前例で「昨年診療を受けた世帯員の割合」の方がより典型的である。
 - » 集落内に正の相関関係があるのが普通であるから。
 - その数量的な表現は教科書 pp. 145-149 に説明してある。

多段抽出法

- 集落抽出法
 - 抽出された集落内の(統計)単位を全て調査する。
 - もし、集落内の単位の数膨大な場合には？
- 二段抽出法
 - (一段目として)抽出された集落において、
(二段目として)(統計)単位を無作為抽出する。
- 多段抽出法
 - 二段抽出を一般化したもの。

多段抽出法

- 二段抽出法のイメージ



多段抽出法

- 例

- － 総務省「家計調査」

- 一段目

- － 市町村を無作為抽出する。

- 二段目

- － (一段目で抽出された)市町村内の調査区を無作為抽出する。

- 三段目

- － (二段目で抽出された)調査区内の世帯を無作為抽出する。

多段抽出法

- 長所

- 一段目の抽出においては、最終抽出単位の抽出枠（名簿）がなくても実行できる。
 - 最終段の抽出の際には抽出枠が必要になる。
- 集落を地域とすれば、調査員の移動量を少なくできる。
 - 調査にかかる時間・費用を節約できる。

- 短所

- 一般にSRSよりも推定精度が低くなる。

多段抽出法

- 世帯を対象とした公的統計
 - 多段抽出が利用される。
 - 理由
 - 一段目に標本と抽出するときに、全世帯の名簿がない。
 - 国勢調査の調査区を利用することが多い。
- 事業所・企業を対象とした公的統計
 - 一段抽出が利用される。
 - 理由
 - 総務省・経済産業省「経済センサス-基礎調査」に基づく事業所母集団DBが名簿として利用できる。