Sample Surveys 1

MAPS project statistical training

Outline

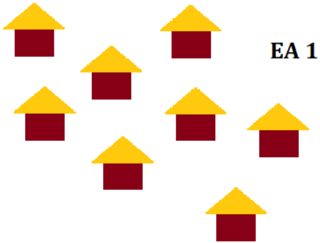
- Introduction to sample surveys: challenges and constraints
- Two-stage cluster sampling: basic ideas
- Notation
- Inclusion probabilities and sample weights
- A general estimator
- Model-based analysis

Introduction

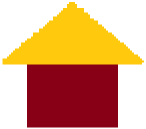
- Sample surveys: typically complex and large-scale tasks such as national-scale surveys of households
- Logistically challenging and costly: sampling a random selection of households from across the country (travel, sensitization, ethics)

Two-stage cluster sampling

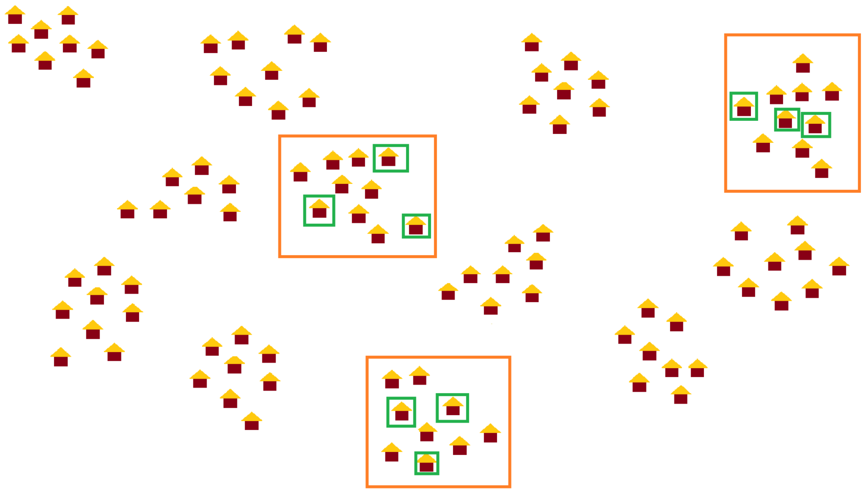
Primary sample units (PSU) e.g. survey enumeration areas (EA)



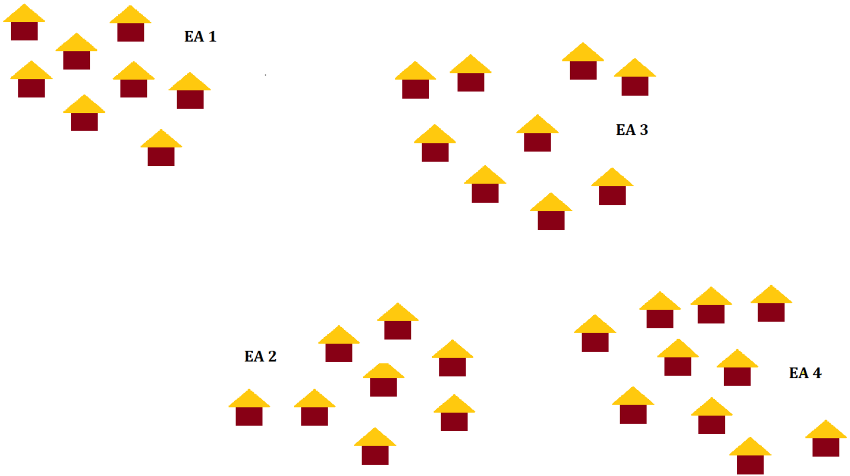
Secondary sample units (SSU) household (HH) within an EA



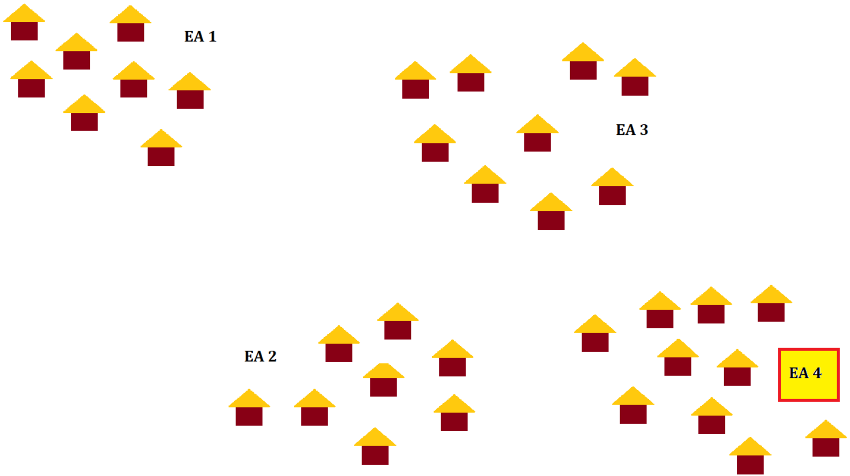
Two-stage cluster sampling



Part of the sampling frame

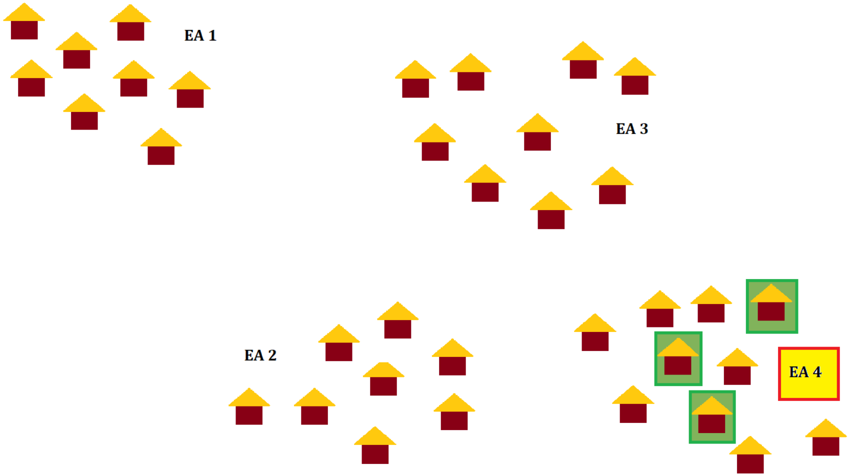


Step 1: selection of EA



Step 2: sensitization

Step 3: sampling SSU within the PSU (here HH within the EA)



Step 4: data collection from selected SSUs

Notation

- N PSUs (EA in our example) in the population..
- We sample n PSUs
- Within each selected PSU we sample SSU, here m_s per PSU
- So the total sample size is $m = n \times m_s$
- m_s might vary from PSU to PSU, m_i in the i^{th} PSU, $m = \sum_{i=1}^{n} m_i$
- The total number of SSU (HH here) in the population is M.

Inclusion probability

- In a design-based sample we can state in advance the inclusion probability for any member of the population
- In a population of N members the *selection* probability for unit i under simple random sampling is $p_i = 1/N$
- If we select a sample of size n with replacement then the probability that unit i is *not* included is $(1 - p_i)^n$
- So the inclusion probability for unit i is

$$\pi_i = 1 - (1 - p_i)^n$$

- If we select a sample of size n without replacement, then the probability that unit i is included in the sample is n/N

Inclusion probability

- If we sample a population of EA by a suitable method we can compute the inclusion probability for the i^{th} EA, π_i^{EA}.
- If the j^{th} HH in the population occurs in i^{th} EA, then the probability that it is selected in a sample from that EA can be computed: $\pi_{j,i}^{\text{HH}}$
- The overall inclusion probability for HH j in a sample from the population is the product $\pi_j = \pi_i^{\text{EA}} \times \pi_{j,i}^{\text{HH}}$
- We sometimes need the joint inclusion probability for two SSU within the population, $\pi_{k,l}$. This is sometimes difficult to obtain.

Estimation of the population total

- If we have a sample of m units, which take values y_i of our target variable and which have inclusion probabilities $\pi_i, i = 1, 2, \ldots m$ then the Horvitz-Thompson (HT) estimate of the population total is given by

$$\widehat{\tau_{\mathrm{HT}}} = \sum_{i=1}^{m} \frac{y_i}{\pi_i}. \qquad (1)$$

Variance of the estimate of the population total

- We may produce a set of estimates of the population total from each sample unit:

$$t_i = \frac{m y_i}{\pi_i},$$

- ... with sample variance

$$s_t^2 = \frac{1}{m-1} \sum_{i=1}^{m} (t_i - \widehat{\tau}_{\mathrm{HT}})^2 .$$

Variance of the estimate of the population total

- A sample variance of the HT estimate of the population total is then given by

$$\operatorname{Var}\left(\widehat{\tau}_{\mathrm{HT}}\right) = \left(\frac{M-m}{M}\right)\frac{s_t^2}{m}. \tag{2}$$

Estimate and standard error of the population mean

- From the estimate of the population total and its variance we can obtain an estimate for the population mean:

$$\widehat{\mu}_{\mathrm{HT}} = \frac{\widehat{\tau}_{\mathrm{HT}}}{M} \tag{3}$$

- ... and its standard error

$$\mathrm{SE}\left(\widehat{\mu}_{\mathrm{HT}}\right) = \frac{\sqrt{\mathrm{Var}\left(\widehat{\tau}_{\mathrm{HT}}\right)}}{M}. \tag{4}$$

Nested random effects model

$$z_{i,j,k,l} \;=\; \mu + \eta_i^{\mathrm{EA}} + \eta_{i,j}^{\mathrm{HH}} + \varepsilon_{i,j,k,l}$$

μ is the mean (constant fixed effect), η^{EA} is a random effect with mean zero and variance σ_{A}^2, for the difference between EAs, and so on for the other random effects. The residual variance component for $\varepsilon_{i,j,k,l}$ is the between-individual within-HH component, but also includes independent measurement error.

Estimation

In a *balanced* hierarchical design the number of units at level m within each unit at level $m-1$ is the same (i.e. the same number of HH in each EA). In this case a simple analysis of variance can be used to estimate variance components.

When a design is unbalanced (deliberately, or by some loss of data), estimation by residual maximum likelihood is preferred

Costs of two-stage sampling

$$C = C_o + nC_{\mathrm{PSU}} + nm_s C_{\mathrm{SSU}}$$

where C_o is fixed overheads costs, C_{PSU} is the cost per PSU and C_{SSU} is the cost per SSU.

Costs of two-stage sampling

C_{PSU}

C_{SSU}

Costs of two-stage sampling

With a fixed budget the optimal value of m_s can be found (assuming this to be fixed over PSU):

$$\tilde{m}_s = \sqrt{\frac{C_{\mathrm{PSU}}\sigma_{\mathrm{w}}^2}{C_{\mathrm{SSU}}\left(\sigma_{\mathrm{b}}^2 - \sigma_{\mathrm{w}}^2/\bar{M}\right)}} \tag{5}$$

where $\bar{M}$ is the number of SSU (HH) in each PSU (EA), assumed to be uniform.

Costs of two-stage sampling

If the budget is fixed at B then:

$$\tilde{n} = \frac{B - C_{\text{o}}}{C_{\text{PSU}} + \tilde{m}_s C_{\text{SSU}}}. \tag{6}$$