Unequal Probability Sampling: Examples

Survey Sampling
Statistics 4234/5234
Fall 2018

November 13, 2018

Two-stage cluster sampling

Possibly unequal selection probabilities

Sample n of the N clusters, without replacement, using selection probabilities ψ_i. Denote the sample by $\mathcal{S}$.

Second stage is an SRS, m_i of the M_i ssus for $i \in \mathcal{S}$.

The probabilities

$$P(\text{cluster } i \text{ included in sample}) = \pi_i$$

are not so easy to evaluate.

The joint probabilities

$$P(\text{clusters } i \text{ and } k \text{ both included}) = \pi_{ik}$$

are even harder, and there are $\binom{N}{2}$ of them!

If the π_i's and π_{ik}'s are too hard, do something:

- *Approximate*: In $\hat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{\hat{t}_i}{\pi_i}$, use the "with-replacement approximation" $\pi_i \approx n\psi_i$ and report

$$\hat{t}_{\text{HT,mod}} = \frac{1}{n} \frac{\hat{t}_i}{\psi_i}$$

- *Conservative*: Another with-replacement approximation

$$\hat{V}_{\text{WR}}(\hat{t}_{\text{HT,mod}}) = \frac{1}{n} \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{\hat{t}_i}{\psi_i} - \hat{t} \right)^2$$

Sampling Weights

(6.4.4)

1. The sampling weight for the ith psu.

   Well

$$\widehat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \frac{\widehat{t}_i}{\pi_i} = \sum_{i \in \mathcal{S}} w_i \widehat{t}_i$$

   so

$$w_i = \frac{1}{\pi_i}$$

   of course.

2. The sampling weight for the jth ssu in the ith psu:

   We have

   $$\pi_{j|i} = P\,[\text{ssu } (i,j) \text{ is in sample} \mid \text{psu } i \text{ is in sample}] = \frac{m_i}{M_i}$$

   assuming SRS at the second stage.

   Then

   $$\widehat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{\pi_{j|i}}$$

   and thus

   $$\widehat{t}_{\mathsf{HT}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{\pi_i \pi_{j|i}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

   with

   $$w_{ij} = \frac{1}{\pi_i \pi_{j|i}} = \frac{1}{P\,[\text{unit } (i,j) \text{ included in sample}]}$$

The Horvitz-Thompson estimator of the population mean is

$$\hat{\bar{y}}_{\mathsf{HT}} = \frac{\hat{t}_{\mathsf{HT}}}{\hat{M}_0} = \frac{\sum_{i \in \mathcal{S}} \frac{\hat{t}_i}{\pi_i}}{\sum_{i \in \mathcal{S}} \frac{M_i}{\pi_i}}$$

The estimated variance is found using theory from ratio estimation (Chapter 4). See equation (6.34) on page 247 for the final result.

PPP sampling

Example: Estimate the total volume of timber in a region.

For each of N trees, let y_i = timber volume. Let x_i = estimated timber volume based on visual inspection.

Logical to take $\psi_i \propto x_i$. Hence the name **probability proportional to prediction**.

Sample will disproportionately include larger trees, which is good.

Lohr (pages 251–252) describes a variation of this approach that requires only one pass through the first. It is an example of **Poisson sampling** since the resulting sample size is random, not fixed.

Dollar unit sampling

Let x_i = book value and y_i = audited value, for $i = 1, 2, \ldots, N$.

The population here is N assets in a portfolio.

Ratio estimation would be a reasonable approach to this problem.

So would be stratified sampling. Stratify by book value, and sample a higher proportion of assets in the strata corresponding to bigger values of x_i.

A third approach is unequal probability sampling with $\psi_i \propto x_i$. This is called **dollar unit sampling**.

Each dollar of book value has the same chance of being included in the sample.