Unequal Probability Sampling: One-Stage with Replacement

Survey Sampling

Statistics 4234/5234

Fall 2018

October 30, 2018

Example (page 219): Consider the population of nursing home residents in a large metropolitan area. Suppose there are 294 such homes, with a total of 37,652 beds.

A two-stage cluster sample (with equal selection probabilities) would take an SRS of nursing homes, then another SRS of residents within each selected home.

In a cluster sample with equal probabilities, however, a nursing home with 20 beds is as likely to be chosen for the sample as a nursing home with 1000 beds.

This approach (sampling homes with equal probabilities) has three major shortcomings:

1. Because we expect the t_i to be proportional to the M_i, the estimators of t and $\bar{y}_U$ may have large variance.

2. A self-weighting equal-probability sample ($m_i \propto M_i$) may be cumbersome to administer, e.g., may require visiting a home just to interview one or two residents.

3. The cost of the sample is unknown in advance.

Instead of taking a cluster sample of homes with equal probabilities, the investigators randomly drew a sample of 57 nursing homes with probabilities proportional to the number of beds. Then they took an SRS of 30 occupants from each sampled home.

Assuming a perfect alignment between *beds* and *occupants* (which of course there isn't), every resident has the same probability of being included in the sample.

The cost is known before selecting the sample, and the estimator of a population total will likely have a smaller variance.

Unequal-probability sampling

We can use unequal inclusion probabilities to decrease variance.

In unequal-probability sampling, we deliberately vary the probabilities of selecting different psus for the sample, then compensate by providing suitable weights in the estimation.

Notation:

$$P(\text{unit } i \text{ selected on first draw}) = \psi_i$$

and

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

You might think sampling with unequal probabilities results in *selection bias*. But because these probabilities are *known*, we can compensate for the unequal probabilities in the weighting.

Sampling only one psu

Suppose we wish to estimate the population total $t = \sum_{i=1}^{N} t_i$ based on observation of one randomly selected (not necessarily with probability $1/N$) psu total t_i.

Then $\pi_i = \psi_i = P(\text{psu } i \text{ is selected})$ for $i = 1, 2, \ldots, N$.

We compensate for the unequal probability of selection by also using ψ_i in the estimator, so

$$w_i = \frac{1}{P(\text{unit } i \text{ is sample})} = \frac{1}{\psi_i}$$

and the estimator is

$$\hat{t}_\psi = \sum_{i \in \mathcal{S}} w_i t_i = \sum_{i \in \mathcal{S}} \frac{t_i}{\psi_i}$$

Properties of estimator:

- Of course $\widehat{t}_\psi$ is unbiased, since

$$E(\widehat{t}_\psi) = \sum_{\mathcal{S}} \widehat{t}_\psi(\mathcal{S}) P(\mathcal{S}) = \sum_{i=1}^{N} \frac{t_i}{\psi_i}\psi_i = \sum_{i=1}^{N} t_i = t$$

- The variance of $\widehat{t}_\psi$ is

$$\begin{aligned} V(\widehat{t}_\psi) &= E\left[(\widehat{t}_\psi - t)^2\right] \\ &= \sum_{\mathcal{S}} \left[\widehat{t}_\psi(\mathcal{S}) - t\right]^2 P(\mathcal{S}) \\ &= \sum_{i=1}^{N} \left(\frac{t_i}{\psi_i} - t\right)^2 \psi_i \end{aligned}$$

Unequal-probability sampling is more efficient than equal probability sampling if the selection probabilities are positively correlated to the cluster totals.

If they are perfectly correlated, $\psi_i \propto t_i$, then $V(\hat{t}_\psi) = 0$, even for $n = 1$.

One-stage sampling with replacement

We will continue to let

$$\psi_i = P(\text{select unit } i \text{ on first draw})$$

for each $i = 1, 2, \ldots, N$.

The idea here is to estimate the population total for each psu drawn, that is

$$\left\{ \frac{t_i}{\psi_i} : i \in \mathcal{R} \right\}$$

represent n independent estimates of t. (Some of which may be duplicates; in that case we keep them all.)

Estimate the population total by the average of the n independent estimates.

The variance of this estimator is

$$\frac{1}{n} \times (\text{the variance for } n = 1)$$

Letting $\mathcal{R}$ denote our sample without replacement of n out of N psus, with selection probabilities $(\psi_1, \psi_2, \ldots, \psi_N)$:

The estimator described above comes down to

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{t_i}{\psi_i}$$

and its variance

$$V(\hat{t}_\psi) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{t_i}{\psi_i} - t \right)^2 \psi_i$$

is estimated by

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n} \frac{1}{n-1} \sum_{i \in \mathcal{R}} \left(\frac{t_i}{\psi_i} - \hat{t} \right)^2$$

Estimating the population mean

We estimate

$$\bar{y}_U = \frac{t}{M_0} = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i}$$

by ratio estimation

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\widehat{M}_{0\psi}} = \frac{\sum_{i \in \mathcal{R}} \frac{t_i}{\psi_i}}{\sum_{i \in \mathcal{R}} \frac{M_i}{\psi_i}}$$

Applying the results from Chapter 4 on ratio estimation:

We have

$$\hat{V}\left(\hat{\bar{y}}_{\psi}\right) = \frac{s_r^2}{n\hat{M}_{0\psi}} \quad \text{(no fpc!)}$$

where

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \hat{\bar{y}}_{\psi}\frac{M_i}{\psi_i}\right)^2$$

the sample variance of $\left\{\frac{t_i}{\psi_i} - \hat{\bar{y}}_{\psi}\frac{M_i}{\psi_i} : i \in \mathcal{R}\right\}$.