Stat 440, E. Slud Oct. 27, 2005

Biased Estimation of S_t^2 in 2-Stage Cluster Sampling

The object of this handout is two-fold:

(i) to show that the naive estimator s_t^2 of the population variance S_t^2 of PSU totals t_i, $i = 1, \ldots, N$ is **biased**; but

(ii) to show that nevertheless, the standard estimator (5.25) in Lohr's book for the theoretical variance $V(\hat{t}_{unb})$ of the unbiased estimator of the population total in a 2-stage cluster SRS is **unbiased**.

Notations are as in Chapter 5 of Lohr's book: the first-stage SRS sample $\mathcal{S}$ consists of n PSU's out of the total of N, and the second-stage SRS takes m_i out of M_i SSU's within each of the first-stage sampled PSU's $i \in \mathcal{S}$. Recall that the PSU totals of attribute values Y_{ij} associated with the j'th SSU in the i'th PSU are denoted $t_i = \sum_{j=1}^{M_i} Y_{ij}$, with population total $t = \sum_{i=1}^{N} t_i$.

Part (i). The naive estimator of $S_t^2 = (N-1)^{-1} \sum_{i=1}^{N} (t_i - t/N)^2$ based on sampled data is

$$s_t^2 \;=\; (n-1)^{-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - n^{-1} \sum_{k \in \mathcal{S}} t_k)^2$$

We calculate the expectation, first all by evaluating conditionally given $\mathcal{S}$ the expected squares as variances plus squares of expectations :

$$E(s_t^2) \;=\; E\left[E\left(\frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i \,-\, \frac{1}{n} \sum_{k \in \mathcal{S}} \hat{t}_k)^2 \,\Big|\, \mathcal{S} \right) \right]$$

$$= \; E\left(\frac{1}{n-1} \sum_{i \in \mathcal{S}} (t_i \,-\, \frac{1}{n} \sum_{k \in \mathcal{S}} t_k)^2 \right) \,+\, E\left[\frac{1}{n-1} \sum_{i \in \mathcal{S}} \mathrm{Var}(\hat{t}_i \,-\, \frac{1}{n} \sum_{k \in \mathcal{S}} \hat{t}_k \,\Big|\, \mathcal{S}) \right]$$

The first expression in the last line is just S_t^2 because sample variances unbiasedly estimate population variances in single-stage SRS sampling; and the second-term integrand for fixed $\mathcal{S}$ is given by standard SRS variance formulas for the variances of the independent second-stage estimators $\hat{t}_i$, using the identity

$$\hat{t}_i \,-\, n^{-1} \sum_{k \in \mathcal{S}} \hat{t}_k \;=\; \frac{n-1}{n}\, \hat{t}_i \,+\, \frac{1}{n} \sum_{k \in \mathcal{S}: k \neq i} \hat{t}_k$$

So $E(s_t^2) \,-\, S_t^2 \,=$

$$E\left[\frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{M_i^2}{m_i} (1 - \frac{m_i}{M_i}) (\frac{n-1}{n})^2 S_i^2 \,+\, \frac{1}{n^2} \sum_{k \in \mathcal{S}: k \neq i} \frac{M_k^2}{m_k} (1 - \frac{m_k}{M_k})\, S_k^2 \right) \right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{N}\frac{n}{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})(\frac{n-1}{n})^2 S_i^2 + \frac{1}{n-1}\sum_{k=1}^{N}\frac{1}{n^2}\frac{(n-1)n}{N}\frac{M_k^2}{m_k}(1-\frac{m_k}{M_k})S_k^2$$

$$= \sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})S_i^2\left\{\frac{n-1}{nN}+\frac{1}{nN}\right\} = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})S_i^2$$

Thus we have proved s_t is a positively biased estimator of S_t^2, with

$$E(s_t^2) = S_t^2 + \frac{1}{N}\sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})S_i^2 \qquad (*)$$

Part (ii). Now recall that the theoretical variance formula – numbered (5.22) in Lohr's book – for $\hat{t}_{unb} = (N/n)\sum_{i\in\mathcal{S}}\hat{t}_i$ is

$$\mathrm{Var}(\hat{t}_{unb}) = N^2(1-\frac{n}{N})\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}S_i^2$$

and recall also that the second-stage sample variances s_i^2 unbiasedly estimate the corresponding PSU attribute-variances S_i^2 given $i\in\mathcal{S}$. We show now that

$$\hat{V}(\hat{t}_{unb}) = N^2(1-\frac{n}{N})\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in\mathcal{S}}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}s_i^2$$

is an unbiased estimator of $\mathrm{Var}(\hat{t}_{unb})$. According to formula (*) and the unbiasedness of s_i^2 for S_i^2 given $i\in\mathcal{S}$, we have $E(\hat{V}(\hat{t}_{unb})) =$

$$\frac{N^2}{n}(1-\frac{n}{N})\left(S_t^2+\frac{1}{N}\sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})S_i^2\right) + E\left[\frac{N}{n}\sum_{i\in\mathcal{S}}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}S_i^2\right]$$

$$= \frac{N^2}{n}(1-\frac{n}{N})\left(S_t^2+\frac{1}{N}\sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-\frac{m_i}{M_i})S_i^2\right) + \sum_{i=1}^{N}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}S_i^2$$

$$= \frac{N^2}{n}(1-\frac{n}{N})S_t^2 + \left(\frac{N}{n}-1+1\right)\sum_{i=1}^{N}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}S_i^2$$

which agrees, as asserted, with the formula for $\mathrm{Var}(\hat{t}_{unb})$ given above.