Survey Sampling: Introduction 2

Cluster Sampling:

One stage cluster sampling:

—Example: Sampling students in high school.

- Take a random sample of classes (The classes are the primary sampling units (psus) or clusters)
- Then measure all students in the selected classes (The students within the classes are the secondary sampling units (ssus))
- Often the ssus are the elements of the population.
- In design of experiments, we would call this a nested design

Comparison with stratification and SRS

- We partition the population into subgroups (strata or clusters)
- With stratification, we sample from each of the subgroups
- With cluster sampling, we sample all of the units in a subset of subgroups
- In general, for a given total sample size n, Cluster sampling will produce estimates with the largest variance. SRS will be intermediate. Stratification will give the smallest variance.

Notation

——PSU level

- N = number of psus in the population
- M_i = number of ssus in the ith psu
- $K = \sum_{i=1}^{N} M_i$= total number of ssus in the population
- y_{ij} = Measurement for jth element in the ith psu
- $t_i = \sum_{j=1}^{M_i} y_{ij}$= total in the ith psu.

- $t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$= population total.

- $S_t^2 = \sum_{i=1}^{N} \frac{(t_i - \frac{t}{N})^2}{N-1}$ = population variance of the psu totals (between cluster variation).

——SSU level

- $\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{K}$ = population mean

- $\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$ = population mean in the ith psu

- $S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{K-1}$ = population variance (per ssu)

- $S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1}$ = population variance within the ith psu.

——Sample values

- n= number of psus in the sample
- m_i = number of elements in the sample for the ith psu
- $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i}$ = sample mean (per ssu) for ith psu
- $\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij}$ = estimated total for the ith psu
- $\hat{t}_{unb} = \sum_{i \in S} \frac{N}{n} \hat{t}_i$ = unbiased estimator of t (population total) (weighted mean of t_i's)

- $s_t^2 = \frac{1}{n-1}\sum_{i\in S}(\hat{t}_i - \frac{\hat{t}_{unb}}{N})^2$= estimated variance of psu totals

- $s_i^2 = \sum_{j\in S_i}\frac{(y_{ij}-\bar{y}_i)^2}{m_i-1}$ = sample variance within the ith psu

Clusters of equal sizes:

$$\hat{t} = \frac{N}{n} \sum_{i \in S} t_i$$

$$V(\hat{t}) = N^2 (1 - \frac{n}{N}) \frac{S_t^2}{n}$$

S_t^2 is estimated by s_t^2 with $s_t^2 = \dfrac{1}{n-1} \sum_{i \in S} (t_i - \dfrac{\hat{t}}{N})^2$

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}$$

$$V(\hat{\bar{y}}) = (1 - \frac{n}{N}) \frac{s_t^2}{nM^2}$$

Source	df	Sum of Squares	Mean Squares
Between psu's	$N-1$	SSB= $\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU}-\bar{y}_U)^2$	MSB
Within psu's	$N(M-1)$	SSW= $\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2$	MSW
Total	NM-1	SSTO= $\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2$	S^2

Example: A student wants to estimate the average grade point average (GPA) in his dormitory. Instead of obtaining a listing of all students in the dorm and conducting a simple random sample, he notices that the dorm consist of 100 suits, each with 4 students; he chooses 5 of those suites at random, and asks every person in the 5 suits what her or his GPA is. The results are as follows:

Person	suit1	suit2	suit3	suit4	suit5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08

The psu's are the suits, so $N = 100$, $n = 5$, and $M = 4$.

$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + 8.96 + 12.96 + 11.08) = 1130.4$$

and

$$
\begin{aligned}
s_t^2 &= \frac{1}{5-1}[(12.16-11.304)^2+\cdots+(11.08-11.304)^2 \\
&= 2.256
\end{aligned}
$$

$$\hat{V}(\hat{t}) = 65.4706$$

$$\hat{\bar{y}} = 1130.4/400 = 2.826$$

$$SE(\hat{\bar{y}}) = \sqrt{(1-\frac{5}{100})\frac{2.256}{(5)(4)^2}} = .164$$

Note: Only the "total" column of the data table is used, the individual GPAs are only used for their contribution to the suite total.

ANOVA Table

Source	df	SS	MS
Between Suites	4	2.2557	.56392
Within suites	15	2.7756	.18504
Total	19	5.0313	.2648

Weight: One-stage cluster sampling with an SRS of psu's produces a self-weighting sample. The weight for each observation unit is

$$w_{ij} = \frac{1}{P\{\text{ssu} \quad j \quad \text{of psu} \quad i \quad \text{is in sample}\}} = \frac{N}{n}$$

$$\begin{aligned}
\hat{t} &= \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij} \\
&= \frac{N}{n}(3.08 + 2.60 + \cdots_{+} 3.28 + 3.20) \\
&= \frac{100}{5}(56.52) \\
&= 1130.4
\end{aligned}$$

Two-stage cluster sampling

- If the items within a cluster are very similar, no need to measure all of them. Alternative is to take an SRS of the units in each selected psu (cluster).

- First: take an SRS of n psus from the population (N psus). Second: For each of the sampled clusters, draw an SRS of size m_i.

- Need to estimate t_i. The sample mean for cluster i is

$$\bar{y}_i = \frac{1}{m_i} \sum_{j \in \text{cluster} i} y_{ij}$$

To estimate the total for cluster i we multiply by M_i, $\hat{t}_i = M_i \bar{y}_i$.

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} \hat{t}_i = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i$$

Estimated variance

- The estimated variance for $\hat{t}_{unb}$ is obtained by deriving a formula for the true variance and substituting sample estimates for unknown parameters in the formula.

- Variance contains two terms: A term equal to the expression for one-stage clustering (S_t^2). An additional term to account for the fact that we took an SRS at the second stage (S_i^2's). The derivation is given in the text for the general case of unequal probability sampling in Section 6.6.

Between cluster variance

- Viewing the $\hat{t}_i$ as an SRS

$$s_t^2 = \sum_{i \in sample} (\hat{t}_i - \hat{\bar{t}})^2/(n-1)$$

where $\hat{\bar{t}} = \hat{t}_{unb}/N$

- s_t^2 is an estimate of S_t^2 the true variance of the t_i

Within cluster variance

- viewing the y_{ij} as an SRS.
- For cluster i, $s_i^2 = \dfrac{1}{m_i - 1} \sum_{j \in psui} (y_{ij} - \bar{y}_i)^2$
- fpc for each cluster, $fpc_i = (1 - m_i/M_i)$
- $\hat{V}(\hat{t}_{unb}) = N^2(1 - \dfrac{n}{N})\dfrac{s_t^2}{n} + \dfrac{N}{n}\sum_{i \in S}(1 - \dfrac{m_i}{M_i})M_i^2\dfrac{s_i^2}{m_i}$

Example 6.1

- Survey of nursing home residents in Philadelphia to determine preferences on life- sustaining treatments
- 294 nursing homes with a total of 37,652 beds (number of residents not known at the planning stage)
- Use cluster sampling
- Suppose we choose an SRS of the 294 nursing homes and then an SRS of 10 residents of each selected home
- A nursing home with 20 beds has the same probability of being sampled as a nursing home with 1000 beds
- 10 residents from the 20 bed home represent fewer people than 10 residents from 1000 bed home

Possible design?

- The above procedure gives a sample that is not self-weighted
- A one-stage cluster sample
- Sample a fixed percentage of the residents of each selected nursing home
- Two-stage cluster design (SRS of homes, then equal proportion SRS of residents in each selected home)
- SRS at first stage, we would expect t_i to be proportional to the number of beds in nursing home i, so estimators will have large variance

The study

- They drew a sample of 57 nursing homes with probabilities proportional to the number of beds
- Then took an SRS of 30 beds (and their occupants) from a list of all beds within each selected nursing home.
- Each bed is equally likely to be in the sample (note beds vs occupants)
- The cost is known before selecting the sample
- The same number of interviews is taken at each nursing home
- The estimators will have smaller variance

Unequal probabilities

- π_i is the probability that unit i is selected as part of the sample
- Most designs we have studied so far have the π_i equal
- In general designs, π_i can vary with i
- Unequal probability sampling may give much better results
- We compensate unequal probabilities by using weights in the estimation

One stage sampling with replacement (suppose $n > 1$)

- $\psi_i = p($select unit i on first draw$)$
- Probability that item i is selected on the first draw is the same as the probability that item i is selected on any other draw
- $\pi_i = p($unit i in sample$)$
- This implies $\pi_i = 1 - (1 - \psi_i)^n$
- $Q_i =$ number of times unit(psu) i occurs in the sample
- $\sum_{i=1}^{N} Q_i = n$, $E(Q_i) = n\psi_i$

- Estimator of total $\hat{t}_\psi = \frac{1}{n} \sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}$
- Sampling with replacement gives us n independent estimates of the population total, one for each unit in sample.
- We average these n estimates

- Unbiased

$$
\begin{aligned}
E(\hat{t}_\psi) &= \frac{1}{n}\sum_{i=1}^{N} E(Q_i)\frac{t_i}{\psi_i} \\
&= \frac{1}{n}\sum_{i=1}^{N} n\psi_i\frac{t_i}{\psi_i} \\
&= \sum t_i \\
&= t
\end{aligned}
$$

Variance:

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

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

$$E[\hat{V}(\hat{t}_\psi)] = V(\hat{t}_\psi)$$

Two-stage sampling with replacement

- The only difference between two-stage sampling with replacement and one-stage sampling with replacement is that in two-stage sampling, we must estimate t_i.
- If psu i is in the sample more than once, there are Q_i estimates of the total for psu i: $\hat{t}_{i1}, \hat{t}_{i2}, \cdots, \hat{t}_{iQ_i}$

- $\hat{t}_\psi = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}$

- $\hat{V}(\hat{t}_\psi) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{(\frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi)^2}{n-1}$

Unequal probability sampling without replacement

- $\pi_i = p(\text{unit} \quad i \quad \text{in sample})$
- π_i/n is the average probability that a unit will be selected on one of the draws: It is the probability we would assign to the ith unit's being selected on draw $k(k = 1, \cdots, n)$ if we did not know the true probabilities
- the estimator $\hat{t}_i/\psi_i$ is then estimated by $\hat{t}_i/(\pi_i/n)$

Horvitz-Thompson (HT) Estimator (Horvitz and Thompson 1952)

$$
\begin{aligned}
\hat{t}_{HT} &= \frac{1}{n}\sum_{i\in S}\frac{\hat{t}_i}{\pi_i/n} \\
&= \sum_{i\in S}\frac{\hat{t}_i}{\pi_i} \\
&= \sum_{i=1}^{N} Z_i\frac{\hat{t}_i}{\pi_i}
\end{aligned}
$$

Unbiased: $E[\hat{t}_{HT}] = \sum_{i=1}^{N}\pi_i\frac{t_i}{\pi_i} = t$

Variance:

$$\hat{V}_1[\hat{t}_{HT}] = \sum_{i\in S}(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i\in S}\sum_{k\in S,k\neq i}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k} + \sum_{i\in S}\frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

Sen-Yates-Grundy form

$$\hat{V}_2[\hat{t}_{HT}] = \sum_{i\in S}\sum_{k\in S,k>i}\frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_k}{\pi_k})^2 + \sum_{i\in S}\frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

Durbin(1953): Use with-replacement variance estimators to avoid some of the potential instability and computational complexity.

In conclusion:

Population total is estimated by

$$\hat{t} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$$

Population mean is estimated by

$$\hat{\bar{y}} = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}$$