STA304/1003 H1 F - Summer 2014:
Surveys, Sampling, and Observational Data

Lecture 9 - Part I:
Two-Stage Cluster Sampling (con'd)

Ramya Thinniyam

June 12, 2014

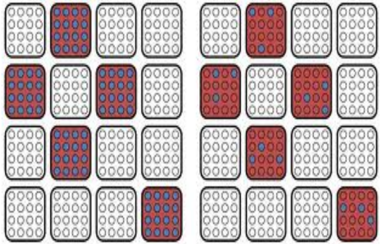
Two-Stage Cluster Sampling

In One-Stage Cluster sampling, all ssus in the selected psus are selected. In Two-Stage Cluster sampling:

1. Select an SRS $\mathcal{S}$ of n psus from the population of N psus.
2. Select an SRS of m_i ssus from each sampled psu i

$\rightarrow$ 2 sources of variability: from selecting psus and selecting ssus (both stages)

Diagram: One-Stage vs. Two-Stage Cluster Samples:



(a) Cluster sampling with $m = 5$ (b) Two-stage sampling with $m = 5$ and $n_j = 3$

Review of Notation

Population Quantities at psu level:

- $N =$ number of psus in the population
- $M_i =$ number of ssus in psu i , $i = 1, 2, \ldots, N$
- $M = \sum_{i=1}^{N} M_i =$ total number of ssus in the population
- $\overline{M} = M/N =$ average cluster size for the population
- $y_{ij} =$ measurement for jth element in psu i
- $\tau_i = \sum_{j=1}^{M_i} y_{ij} =$ total in psu i
- $\tau = \sum_{i=1}^{N} \tau_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} =$ population total
- $S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \frac{\tau}{N})^2 =$ population variance of the psu totals

Population Quantities at ssu level:

- $\bar{y}_U = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} =$ population mean
- $\bar{y}_{iU} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \frac{\tau_i}{M_i}$ population mean in psu i
- $S^2 = \frac{1}{M-1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_U)^2 =$ population variance (per ssu)
- $S_i^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iU})^2 =$ population variance within psu i

Sample Quantities

- $n =$ number of psus in the sample
- $m_i =$ number of ssus in the sample from psu i
- $\mathcal{S}$: sample of psus
- $\mathcal{S}_i$: sample of m_i ssus from ith psu
- $\bar{y}_i = \frac{1}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} =$ sample mean for psu i
- $\hat{\tau}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i =$ estimated total for psu i
- $s_i^2 = \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2 =$ sample variance within psu i

Estimating the Population Mean

1. M is known:

   $\hat{\bar{y}}_{unb} = \frac{N}{M}\sum_{i\in\mathcal{S}}\frac{M_i\bar{y}_i}{n} = \frac{\hat{\tau}_{unb}}{M}$ is an unbiased estimator of the population mean

   - $E(\hat{\bar{y}}_{unb}) = \bar{y}_U$
   - $\hat{V}(\hat{\bar{y}}_{unb}) = \frac{1}{n\bar{M}^2}\left(1-\frac{n}{N}\right)s_b^2 + \frac{1}{nN\bar{M}^2}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$;
     where
     $s_b^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - \overline{M}\hat{\bar{y}}_{unb})^2$ is the sample variance among the $M_i\bar{y}_i$ terms.

2. M is unknown. Use Ratio Estimation:

   $\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}}\hat{\tau}_i}{\sum_{i\in\mathcal{S}}M_i} = \frac{\sum_{i\in\mathcal{S}}M_i\bar{y}_i}{\sum_{i\in\mathcal{S}}M_i}$

   - $\hat{V}(\hat{\bar{y}}_r) = \frac{1}{n\bar{M}^2}\left(1-\frac{n}{N}\right)s_r^2 + \frac{1}{nN\bar{M}^2}\sum_{i\in\mathcal{S}}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$
     When N is large, the second term is negligible compared to first.

   Recall: $s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$

Estimating the Population Total

Unbiased Estimation:

$\hat{\tau}_{unb} = \frac{N}{n}\sum_{i\in\mathcal{S}}\hat{\tau}_i = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i = \frac{N}{n}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{M_i}{m_i}y_{ij}$
is an unbiased estimator of population total

$\hat{\tau}_i$'s are random variables so $\hat{\tau}_{unb}$ has 2 sources of variability:

(1) variability between psus

(2) variability of ssus within psus

Properties of $\hat{\tau}_{unb}$:

- $E(\hat{\tau}_{unb}) = \tau$
- $\hat{V}(\hat{\tau}_{unb}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)s_b^2 + \frac{N}{n}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$
  $\hookrightarrow$ Variance from one-stage cluster + additional variance due to selection of ssus within psus

Design Issues

1. Precision Needed:
   - Determine ME, e

2. Choosing the psu size:
   - Mostly natural like clutches of eggs, classes with students, etc. Sometimes have choice such as area of forest, time interval between costumers.
   - More area $\Rightarrow$ more variability within psus $\Rightarrow$ ICC smaller

3. Choosing subsampling sizes (how many ssus to sample in each psu):
   - Assuming equal cluster sizes, $\overline{M}$ and take equal sample sizes m - minimize variance for fixed cost
   - $V(\hat{\bar{y}}_{unb}) = \left(1-\frac{n}{N}\right)\frac{MSB}{nM} + \left(1-\frac{m}{M}\right)\frac{MSW}{nm}$:

     If $MSW = 0$, $R_a^2 = 1$: choose $m = 1$. For other values, depends on relative costs.
   - total cost = $C = c_1 n + c_2 nm$:
   - $n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$ and $m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$:

     Estimate R_a^2 from pilot survey: $\hat{R}_a^2 = 1 - \frac{\widehat{MSW}}{\hat{S}^2}$ and for large populations

     $m_{opt} = \sqrt{c_1(1-\hat{R}_a^2)/c_2\hat{R}_a^2}$
   - For unequal cluster size use $\bar{M}$ instead of M to determine $\bar{m}$: sample $\bar{m}$ in each psu or allocate so that $\frac{m_i}{M_i}$ is constant

4. Choosing the Sample Size (number of psus, n):

- Determine psu size and subsampling fraction. Decide on desired ME, e
- For equal-sized clusters:

$$V(\hat{\bar{y}}) \leq \frac{1}{n}\left[\frac{MSB}{\overline{M}} + \left(1 - \frac{m}{\overline{M}}\right)\frac{MSW}{m}\right] = \frac{v}{n}$$

- $n = z_{\alpha/2}^2 v/e^2$
- Estimate $v = \left[\frac{MSB}{\overline{M}} + \left(1 - \frac{m}{\overline{M}}\right)\frac{MSW}{m}\right]$ from previous survey or prior knowledge

5. Iterate:

- Above gives the n for required ME
- Modify survey design (add stratification, auxiliary variables, etc.) until cost is within budget.

Example: Creamed Corn

psu = case
ssu = can
$N = 580$ cases
$M_i = 24$ for all i, $\bar{M} = 24$
$\sum_{i=1}^{N} M_i = 580(24) =$ total cans in truck $= 13920$
$\widehat{MSW} = MS$ residuals $= 4.53$

$$\hat{S}^2 = \frac{\widehat{SSTO}}{N\bar{M}-1} = \frac{(N-1)MSB + N(\bar{M}-1)MSW}{N\bar{M}-1}$$

$$= \frac{579(13.60) + 580(23)(4.53)}{13.919} = 4.91$$

$$R_a^2 = 1 - \frac{\widehat{MSW}}{\hat{S}^2} = 1 - \frac{4.53}{4.91} = 0.0774$$

$C_1 = 20$
$C_2 = 8$
$C = 120$ mins (budget = total cost)

An inspector samples cans from a truckload of canned creamed corn to estimate the average number of worm fragments per can. The truck has 580 cases; each case contains 24 cans. It takes 20 minutes to locate and open a case, and 8 minutes to locate and examine each specified can within a case. Assume your budget is 120 minutes. A preliminary study of 12 cases at random subsampling 3 cans from each case yields:

C1: 1 5 7	C7: 5 5 1
C2: 4 2 4	C8: 3 0 2
C3: 0 1 2	C9: 7 3 5
C4: 3 6 6	C10: 3 1 4
C5: 4 9 8	C11: 4 7 9
C6: 0 7 3	C12: 0 0 0

$$m_{opt} = \sqrt{\frac{C_1 \bar{M}(N-1)(1-R_a^2)}{C_2(N\bar{M}-1)R_a^2}} = \sqrt{\frac{20(24)(579)(0.9226)}{8(13919)(0.0774)}}$$

$= 5.45 \to 6$ cans

$$n_{opt} = \frac{C}{C_1 + C_2 c_{opt}} = \frac{120}{20+8(5.45)} = 1.89 \to 2 \text{ cases}$$

total cost $= (2 \text{ cases} \times 20) + (8 \times 6 \times 2) = 136$ mins
"over budget"

So sample 2 cases × 5 cans in each.

How many cans should be examined per case? How many cases?

Using 'R' to get ANOVA Table:

```
> case=rep(seq(1,12,1),each=3)
> case
 [1]  1  1  1  2  2  2  3  3  3  4  4  4  5  5  5  6  6  6
      7  7  7  8  8  8  9  9  9 10 10 10 11 11 11 12 12 12

> case=factor(case)
> case
 [1] 1  1  1  2  2  2  3  3  3  4  4  4  5  5  5  6  6  6
     7  7  7  8  8  8  9  9  9  10 10 10 11 11 11 12 12 12
Levels: 1 2 3 4 5 6 7 8 9 10 11 12

> frag=c(1,5,7,4,2,4,0,1,2,3,6,6,4,9,8,0,7,3,5,5,1,3,0,2,7,3,5,3,1,4,4,7,9,0,0,0)
> frag
 [1] 1 5 7 4 2 4 0 1 2 3 6 6 4 9 8 0 7 3 5 5 1 3 0 2 7 3 5 3 1 4 4 7 9 0 0 0


> model <- lm(frag ~ case)
> anova(model)
Analysis of Variance Table

Response: frag
          Df Sum Sq Mean Sq F value  Pr(>F)
case      11 149.64 13.6035  3.0045 0.01172 *
Residuals 24 108.67  4.5278
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary and Advantages/Disadvantages of Cluster Sampling

- Cluster sampling used commonly in large surveys
- Convenient, easy to access elements by clusters since clusters occur naturally together
- In cluster sampling, want elements to be heterogenous within groups; In STRS, want elements to be homogeneous within groups (opposites)
- If elements within clusters are homogenous, two-stage cluster sampling is better
- One-Stage is a special case of the general Two-Stage Cluster sample (using $M_i = m_i$)
- Cluster sampling usually has larger variance than using SRS for the same sample size
- Cluster sampling can give more precision per dollar if measuring individual elements is much more costly than sampling clusters
- Two types of estimation for population parameters: Unbiased and Ratio estimation
  - If cluster sizes vary greatly, ratio estimation is better to use (smaller variance) and may be an advantage to sample with probabilities proportional to cluster size
  - For equal cluster sizes, both types of estimates are equivalent