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

Lecture 8: Two-Stage Cluster Sampling

Ramya Thinniyam

June 10, 2014

Two-Stage Cluster Sampling

In One-Stage Cluster Sampling, we measure all ssus within selected clusters.

If ssus within a cluster are similar, then we do not want to measure them all as it may cause repetition, waste resources, be expensive.

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)

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

Sampling Weights

For two-stage cluster sampling when subsampling by SRS, we have:

$$w_{ij} = \frac{1}{P(ssu\ j\ of\ psu\ i\ is\ in\ sample)} = \frac{N}{n}\frac{M_i}{m_i}$$

$$W_{ij} = \frac{1}{\pi_{ij}}$$

$\rightarrow$ self-weighting sample when m_i is proportional to M_i.

Proof : $\pi_{ij} = P(\text{jth ssu from ith clusters in sample})$

$= P(i \in S \;\&\; j \in S_i) = P(i \in S)\, P(j \in S_i)$ by indep

$= \frac{n}{N} \leftarrow$ (choose psu by SRS) $\cdot \frac{m_i}{M_i} \leftarrow$ (choose by SRS

M_i ssus in psu$i$

m_i ssus in psu$i$ for sample)

Same interpretation as before: ssu j from psu i represents itself $+\ \frac{NM_i}{nm_i} - 1$ unsampled ssus.

We can write the estimators in terms of the weights as follows:

$$\hat{\tau} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

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
     $s_b^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - \bar{M}\hat{\bar{y}}_{unb})^2$ is the sample variance among the $M_i\bar{y}_i$ terms.

$M_i\bar{y}_i \neq y_i \leftarrow$ sample total for ith cluster

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

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

Example: Using R for Two-Stage Cluster Sampling

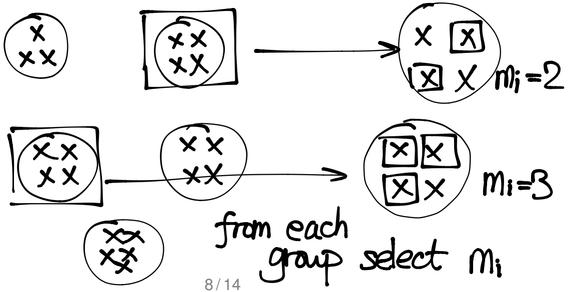
Use ‘Sampling’ Package

```
> algebra <- read.csv("algebra.csv")

> cl = mstage(data=algebra, stage=c("cluster", "stratified"), varnames=c("class","score"),
            size=list(5,c(2,2,3,3,3)),method="srswor")
> twostage=getdata(algebra,cl)
> twostage
[[1]]
    Mi  score  class  ID_unit  Prob_ 1 _stage
53  24    37     38       53        0.4166667
57  24    65     38       57        0.4166667
52  24    60     38       52        0.4166667
.
132 28    65     44      132        0.4166667
133 28    60     44      133        0.4166667
136 28    52     44      136        0.4166667
.
159 19    34     46      159        0.4166667
163 19    71     46      163        0.4166667
160 19    42     46      160        0.4166667
.
220 17   100     58      220        0.4166667
221 17    43     58      221        0.4166667
222 17    48     58      222        0.4166667
.
234 21    71     62      234        0.4166667
227 21    31     62      227        0.4166667
.
.
```

stage=c("cluster","stratified")
(1st stage, 2nd stage)

Ex: $N=5$
$n=2$



```
[[2]]
    class Mi  score  ID_unit  Prob_ 2 _stage        Prob
53     38 24    37       53     0.08333333 0.03472222
68     38 24    68       68     0.08333333 0.03472222
141    44 28    75      141     0.07142857 0.02976190
155    44 28    75      155     0.07142857 0.02976190
160    46 19    42      160     0.15789474 0.06578947
170    46 19    64      170     0.15789474 0.06578947
169    46 19    34      169     0.15789474 0.06578947
220    58 17   100      220     0.17647059 0.07352941
226    58 17    49      226     0.17647059 0.07352941
219    58 17    49      219     0.17647059 0.07352941
247    62 21    61      247     0.14285714 0.05952381
239    62 21    63      239     0.14285714 0.05952381
246    62 21    51      246     0.14285714 0.05952381
```

```
> Mi = tapply(twostage[[1]]$score,twostage[[1]]$class,length)
> Mi
38 44 46 58 62
24 28 19 17 21
```

```
> attach(twostage[[2]])
> mi=tapply(score,class,length)
> mi
38 44 46 58 62
 2  2  3  3  3
> ybari = tapply(score,class,mean)
> ybari
  38    44   46    58     62
52.50 75.00 46.67 66.00 58.33
> ybarhatr = sum(Mi*ybari)/sum(Mi)
> ybarhatr
[1] 60.49235
```

```
> sqi = tapply(score,class,var)
> sqi
  38     44    46     58     62
480.50  0.00 241.33 867.00  41.33
> sum(Mi^2 * (ybari - ybarhatr)^2)
[1] 281630.7
> mean(Mi)
[1] 21.8
> mean(mi)
[1] 2.6
> sum( Mi^2 *(1-mi/Mi)*(sqi/mi) )
[1] 225297.1
```

Example: Two Stage Cluster Sample: Algebra Test Scores by Class

a). psu=class
ssu=student
$N=12 \quad n=5$

sample

class id	38	44	46	58	62
$i\in S$	1	2	3	4	5
M_i	24	28	19	17	21
m_i	2	2	3	3	3
$\bar{y}_i$	52.5	75	46.67	66	58.33
S_i^2	480.5	0	241.2	867	41.33

$\boxed{\hat{\bar{M}}=21.8}$

In a population of 12 algebra classes, an SRS of 5 classes is taken. Then some students in the selected classes are randomly selected and given an algebra test and the scores are recorded. Use the 'R' output (from previous slides) to answer the following:

a) Identify the psu, ssu, N, n, M_i, and m_i.

b) Estimate the mean score in the classes and its standard error.

c) Suppose you are now given the information that there are a total of 299 students in the population. Estimate the mean score and its standard error.

b). Since M (total # students) unknown ⇒ use ratio est.

$$\hat{\bar{y}}_r=\frac{\sum_{i\in S} M_i\bar{y}_i}{\sum_{i\in S} M_i}=\frac{24(52.5)+\cdots+21(58.33)}{24+\cdots+21}=60.49$$

$$Se(\hat{\bar{y}}_r)=\sqrt{\frac{1}{n\bar{M}^2}\left(1-\frac{n}{N}\right)S_r^2+\frac{1}{nN\bar{M}^2}\sum_{i\in S} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i}}$$

$$=\sqrt{\frac{1}{5(21.8)^2}\left(1-\frac{5}{12}\right)\frac{281630.7}{4}+\frac{1}{5\cdot 12(21.8)^2}(225297.1)}$$

$=5.02$

c). $$\hat{\bar{y}}_{unb}=\frac{N}{M}\sum_{i\in S}\frac{M_i\bar{y}_i}{n}$$

$$=\frac{12}{299}\cdot\frac{60.49\times 21.8\times 5}{5}$$

$=52.92$

Example: The Case of the Six-Legged Puppy

a). two-stage cluster sample
$M=40$ puppies
PP: $M_1=30$ PL: $M_2=10$
$n=1$ home selected
$N=2$ homes
$m_i=2$ puppies in total
in sample
b). Puppy Palace is selected
in PP: $\hat{t}_{unb}=30\times4=120$
c). in DL:
$\hat{t}_{unb}=10\times4=40$
$\hat{t}_{unb}=\frac{2}{1}\hat{t}_{PP}=240$
$\hat{y}_{unb}=6$
$\hat{t}_{unb}=\frac{2}{1}\hat{t}_{DL}=2\times40=80$
$\hat{y}_{unb}=2$

We wish to estimate the mean number of legs on healthy puppies in Sample City puppy homes. Sample City has two puppy homes: Puppy Palace with 30 puppies and Dog's Life with 10 puppies. One home is selected randomly and 2 puppies from that home are randomly chosen.

a) What type of sampling method is used here? Identify its parameters ie. the population size, sample size, etc.
b) Suppose Puppy Palace is selected as the home and each sampled puppy has 4 legs. Use $\hat{\bar{y}}_{unb}$ to estimate the mean number of legs per puppy.
c) Suppose Dog's Life is selected as the home and each sampled puppy has 4 legs. Use $\hat{\bar{y}}_{unb}$ to estimate the mean number of legs per puppy.
d) Use Ratio Estimation instead to estimate the mean for the scenarios in b) and c).
e) Why is $\hat{\bar{y}}_{unb}$ a 'bad' estimator even though it is unbiased? Why is ratio estimation better to use in this case?

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:

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

- $n = z^2_{\alpha/2} v / e^2$
- 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

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

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
```