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

Lecture 7: One-Stage Cluster Sampling

Ramya Thinniyam

June 5, 2014

Cluster Sampling

Sampling units are clusters or groups of elements. Elements are sampled from clusters in stages. Elements in a cluster are usually grouped together by location: convenient to sample elements in the same cluster.

- Population is divided into N clusters of elements called primary sampling units (psus)
- n clusters selected using SRS
- Units called secondary sampling units (ssus) are sampled from these clusters (either all units or choose some randomly):
  Psus are usually the elements in the population, but sometimes we can subsample clusters from clusters
- If units in same cluster are similar, less likely to get a representative sample and same information is repeated by sampling all units of a cluster $\rightarrow$ less precision

Why use Cluster Sampling?

- Sampling frame of observation units is not available or difficult to get
- Population occurs in natural clusters. Ex. geographical location
- Economical

When should it be used?

- Elements in a cluster are heterogeneous
- Clusters are homogeneous

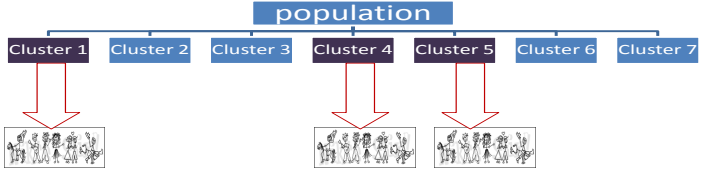
  (opposite of STRS)

  If elements within clusters are homogenous, more information can be obtained by sampling a large number of clusters of smaller size.

  STRS generally increases precision whereas Cluster Sampling decreases it

One-Stage Cluster Sampling

- Population is divided into N clusters of elements (psus)
- n clusters selected using SRS
- All units in selected clusters are measured.
  Ssus are the elements within the clusters that are measured



Examples

- Sample:

(1) Children within schools
(2) Households within city blocks
(3) Workers within factories
(4) Plants within fields

- Primary sampling unit = cluster:

(1) Schools
(2) City blocks
(3) Factories
(4) Fields

- Secondary sampling unit = elements:

(1) Children
(2) Households
(3) Workers
(4) Plants

Difference Between STRS and Cluster Sampling

Identify which is an example of STRS and which is Cluster Sampling:

1. The Dept of Agriculture wants to investigate the use of pesticides by farmers in England. A sample of the different counties in England is chosen at random, and all farmers in the selected counties would be sampled.

Cluster

2. A farmer wishes to work out the average milk yield of each cow type in his herd which consists of Ayrshire, Friesian, Galloway and Jersey cows. He divides his herd into the four types and takes a random sample from each type.

STRS

Notation

Population Quantities at psu level:

- $N =$ number of clusters/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
- $\tau_i = y_i = \sum_{j=1}^{M_i} y_{ij} =$ total in psu i
- $\tau = \sum_{i=1}^{N} \tau_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} =$ population total

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 ,
  $i = 1, 2, \ldots, N$: $m_i = M_i$ for one-stage cluster sampling
- $\mathcal{S}$: sample of psus
- $\mathcal{S}_i$: sample of ssus from ith psu
- $\bar{y}_i = \frac{1}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij} =$ sample mean for psu i
- $\hat{\tau}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = y_i$ estimated total for psu i
- $\bar{y}_t = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i =$ average of the sampled cluster totals
- $s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (y_i - \bar{y}_t)^2 =$ sample variance of psu totals
- $s_i^2 = \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2 =$ sample variance within psu i

Ratio Estimation

Ratio Estimator of Population Mean, $\bar{y}_U$: $\bar{y} = \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i}$;

- Write the population mean as a ratio:

$$\bar{y}_U = \frac{\sum_{i=1}^{N} \tau_i}{\sum_{i=1}^{N} M_i} = \frac{\tau}{M}$$

- Expect y_i to be positively correlated with m_i so we can consider this as a ratio estimation problem with auxiliary variable $x_i = m_i$
- Define:

$$r = \bar{y} = \frac{\hat{\tau}}{\hat{M}} = \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i}$$

- Denominator of estimator for $\bar{y}$ is random: depends on which clusters are chosen for the sample
- $\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n\bar{M}^2}$;
  where $s_r^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (y_i - \bar{y} m_i)^2$
- $\rightarrow$ Variance depends on variability of cluster means
- If M is unknown, use $\bar{m}$ to estimate $\bar{M}$
- If M is known, use Ratio Estimator for Population Total, τ: $\hat{\tau}_r = M\bar{y}$ with $SE(\hat{\tau}_r) = MSE(\bar{y})$
- Ratio estimator of total requires we know the total number of elements in the population, whereas the next estimator $\hat{\tau}_{unb}$ does not

Unbiased Estimation

Unbiased Estimator of Population Total, *tau*:

$\hat{\tau}_{unb} = N\bar{y}_t = \sum_{i \in \mathcal{S}} \frac{N}{n} y_i$

- Does not depend on M (the number of elements in the population)
- $\hat{V}(\hat{\tau}_{unb}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}$; where $s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (y_i - \bar{y}_t)^2$
- If there is large amount of variation among cluster sizes and if cluster sizes are highly correlated with cluster totals, then the ratio estimator will generally have lower variance than the unbiased estimator
- The unbiased estimator does not use the information given by the cluster sizes $m_1, \ldots, m_n$ and so may be less precise

Clusters of Equal Size

- $M_i = m_i = m$
- $M = Nm$ and the total sample size is nm
- Treat the psu means or totals as the observations to do estimation
- SRS of n data $\{y_i, i \in \mathcal{S}\}$
- Estimator of population mean:
  - $\bar{\bar{y}}_c = \frac{1}{mn} \sum_{i \in \mathcal{S}} \sum_{j=1}^{m} y_{ij}$: overall average of all nm sample measurements
  - $\bar{y}_t = m\bar{\bar{y}}_c$
  - $\hat{V}(\bar{\bar{y}}_c) = \left(1 - \frac{n}{N}\right) \frac{1}{nm^2} s_t^2 = \left(1 - \frac{n}{N}\right) \frac{1}{n(n-1)} \sum_{i \in \mathcal{S}} (\bar{y}_i - \bar{\bar{y}}_c)^2$
- Estimator of population total:
  - The ratio estimator ($M\bar{y}$) and unbiased estimator ($N\bar{y}_t$) are equivalent

Example: GPA Estimation

population = all students in the residence at UTM
$M = 400$ psu = house ssu = student → observation unit
$N = 100, n = 5, M_i = m_i = m = 4$ (equal cluster size)
→ treat the total y_i (bottom row of table as data from an SRS)

A student wishes to estimate the mean GPA of the students in a residence at UTM. There are 100 houses each with 4 students. 5 houses are randomly chosen and all students living in those houses are asked to report their GPA.

Student Number	House 1	2	3	4	5
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

$\bar{y}_c = \bar{y}_t / m = 11.304/4 = 2.826$ or $\bar{y}_c = \hat{\tau}_{unb} / NM = 1130.4/100(4) = 2.826$
→ estimates the mean GPA of a student

$S_t^2 = \frac{1}{n}\sum_{i \in S}(y_i - \bar{y}_t)^2 = \frac{1}{4}[(12.16 - 11.304)^2 + \cdots + (11.08 - 11.304)^2] = 2.256$ → estimates the variance of the house totals

$se(\bar{y}_c) = \sqrt{(1 - \frac{n}{N})\frac{S_t^2}{nm^2}} = \sqrt{(1 - \frac{5}{100})\frac{2.256}{5(4)^2}} = 0.1637$

Answer: the mean GPA of a student in this residence is estimated as 2.826 with se of 0.1637.

Find an estimate of the mean GPA in this residence and its standard error.

$\hat{\tau}_{unb} = \frac{N}{n}\sum_{i \in S} y_i = \frac{100}{5}(12.16 + 11.36 + 9.86 + 12.96 + 11.08) = 1130.4$ → estimates the total GPA of all students in this residence.

$\bar{y}_t = \frac{1}{n}\sum_{i \in S} y_i = \frac{1}{5}(12.16 + \cdots + 11.08) = \hat{\tau}_{unb}/100 = 11.304$ → estimate the mean of house totals.

Sampling Weights

For one-stage cluster sampling when choosing psus by SRS, we have:

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

$\rightarrow$ self-weighting sample.

Proof : $P(\text{ssu } i \text{ from psu } i \text{ in sample}) = P(\text{psu } i \text{ is in sample})$, once psu chosen, all ssus in the psu are in sample

$= \frac{n}{N}$ (proved in SRS lec)

N clusters choose n by SRS

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

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

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

Theory for Clusters of Equal Sizes

Cluster sampling almost always gives less precision than an SRS with the same number of elements.

ANOVA Table for the Population:

Source	df	Sum of Squares	Mean Square
Between psus	$N-1$	$SSB = \sum_{i=1}^{N}\sum_{j=1}^{m}(\bar{y}_{iU}-\bar{y}_U)^2$	$MSB = \frac{SSB}{N-1}$
Within psus	$N(m-1)$	$SSW = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{iU})^2$	$MSW = \frac{SSW}{N(m-1)}$
Total (about $\bar{y}_U$)	$Nm-1$	$SSTO = \sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij}-\bar{y}_U)^2 = (Nm-1)S^2$	$MSTO = S^2$

Results:

- $\widehat{MSB}$ and $\widehat{MSW}$ calculated from sample are both unbiased for population MSB and MSW but $\widehat{MSTO}$ is biased for MSTO
- Opposite of STRS - have to worry about unsampled clusters whereas in STRS we sample from each strata
- $V(\bar{\bar{y}}_c) = \left(1-\frac{n}{N}\right)\frac{MSB}{nm}$
- For an SRS of size nm : $V(\bar{y}_{SRS}) = \left(1-\frac{nm}{Nm}\right)\frac{S^2}{nm} = \left(1-\frac{n}{N}\right)\frac{S^2}{nm}$
- If $\frac{MSB}{MSW}$ is large, precision is decreased. If $MSB > S^2$, cluster sampling is less efficient than SRS.

Intraclass Correlation Coefficient

Intraclass/Intracluster Correlation Coefficient (ICC) measures similarity of elements in the same cluster. ie. homogeneity within clusters

$$ICC = 1 - \frac{m}{m-1}\frac{SSW}{SSTO}$$

- $-\frac{1}{m-1} \leq ICC \leq 1$
- $ICC = 1$ when clusters are perfectly homogeneous ie. $SSW = 0$
- $MSB = \frac{Nm-1}{m(N-1)} S^2[1 + (m-1)ICC]$
- $\frac{V(\bar{y}_c)}{V(\bar{y}_{SRS})} = \frac{MSB}{S^2} = \frac{Nm-1}{m(N-1)}[1 + (m-1)ICC]$: ratio of variances / relative efficiency of cluster to SRS estimator
- If N large, then ratio $\approx 1 + (m-1)ICC \rightarrow$ we need to take $[1 + (m-1)ICC]$ as many elements in a cluster sample than we would for a SRS to obtain the same precision
- $ICC > 0$ if $MSB > MSW$: when elements within a psu are similar and so cluster sampling is less efficient than SRS (usually occurs when clusters occur naturally)
- $ICC \leq 0$ if $MSB \leq MSW$: when elements within a psu are dispersed, cluster means will be similar and so cluster sampling more efficient than SRS

Adjusted R^2

ICC is defined for equal cluster sizes.

Another measure of homogeneity is Adjusted R^2, R_a^2 .

$$R_a^2 = 1 - \frac{MSW}{S^2}$$

- Interpreted as proportion of variability in the population (y) explained by the clusters but adjusted for the number of clusters ie. penalize for too many predictors in the regression
- If clusters are homogeneous, then *SSB* high compared to *SSW* and R_a^2 is high
- If all psus are the same size: $\frac{V(\bar{\bar{y}}_c)}{V(\bar{y}_{SRS})} = \frac{MSB}{S^2} = 1 + \frac{N(m-1)}{N-1} R_a^2$: we need to take $[1 + \frac{N(m-1)}{N-1} R_a^2]$ as many elements in a cluster sample than we would for a SRS to obtain the same precision

Example: GPA Estimation

Use the following sample ANOVA table to obtain the SE for the estimate of the mean GPA and also estimate *ICC* and R^2_a. (Hint: the sample MSB and MSW are unbiased estimators of the population quantities, however the sample MSTO is not an unbiased estimator of the population MSTO.)

Source	df	SS	MS
Between Houses	4	2.2557	0.56392
Within Houses	15	2.7756	0.18504
Total	19	5.0313	0.26480

$\hat{M}SB = 0.56392$

$\hat{M}SW = 0.18504$

$\hat{S}^2 = \frac{\hat{SSB} + \hat{SSW}}{Nm - 1} = 0.2790$

$\hat{SSTO} = 111.34$

$\hat{SSB} = (N-1)\hat{M}SB$

$\hat{SSW} = N(m-1)\hat{M}SW$

$E(\hat{M}SB) = MSB$

$E(\hat{M}SW) = MSW$

$N = 100 \quad n = 5 \quad m = 4 \quad M = 400$

$se(\bar{y}_c) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{M}SB}{nm}}$

$\hat{M}SB = 0.56392$ from anova table $= 0.164$ same as before

$\hat{ICC} = 1 - \frac{m}{m-1}\frac{\hat{SSW}}{\hat{SSTO}} = 1 - \frac{4}{3}\left(\frac{55.828}{111.34}\right)$

$\hat{ICC} = 0.335$

$\hat{R}^2_a = 1 - \frac{\hat{M}SW}{\hat{S}^2} = 1 - \frac{0.18504}{0.279} = 0.337$

$\hat{R}^2_a = 0.337$

$\frac{\hat{V}(\bar{y}_c)}{\hat{V}(\bar{y}_{SRS})} = \frac{\hat{M}SB}{\hat{S}^2} = \frac{0.56392}{0.279} = 2.02 = \frac{\hat{V}(\bar{y}_c)}{\hat{V}(\bar{y}_{SRS})}$

Need sample about (2.02n) elements for Cluster sample, but only n for SRS for the same precision.

1 psu $= \frac{4}{2.02} = 1.98$ students using SRS.

Clusters of Unequal Sizes

Difference between equal and unequal cluster sizes:
Variation among individual cluster totals is likely to be large when clusters have different sizes, M_i. ie s_t^2 larger for unequal cluster sizes

- Estimator of mean can be inefficient when M_i are different because it depends on the variability of the cluster totals
- Use ratio estimator $\bar{y} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} m_i}$
- Ratio estimator of population mean is usually more efficient when there are unequal cluster sizes
- Need to know M, not just M_i for sampled clusters to estimate population total (see Ratio Estimation slide)

Example: Algebra Test Scores by Class

$N = 187$ classes
$n = 10$
unequal class sizes

$\bar{M} = \frac{\sum m_i}{n} = 24.7$

sample data:

class id	23	38	39	...	108
i	1	2	3	...	10
m_i (class size)	20	24	34	...	26
y_i (class total)	1230	1402	1972	...	1746

So far cannot say which class is better.

$$\bar{y} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} m_i} = \frac{15417}{247} = 62.417$$

ratio est. for the mean score

OR

$$\bar{y} = \frac{\sum y_i}{\sum m_i} = \frac{\sum m_i \bar{y}_i}{\sum m_i} = 62.417$$

In a population of 187 algebra classes, an SRS of 10 classes is taken. Each student in the selected classes is given an algebra test and the scores are recorded. Use the 'R' output to estimate the mean score in the classes and its standard error.

$$se(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n\bar{M}^2}} = \sqrt{\left(1 - \frac{10}{187}\right)\frac{193470.1}{9(10)(24.7)^2}}$$

where $s_r^2 = \frac{1}{n-1}\sum_{i \in S}(y_i - \bar{y}m_i)^2 = \frac{193470.1}{9}$

Use package 'Sampling':

```
> algebra <- read.csv("algebra.csv")
> algebra
    class Mi score
1       23 20    57
2       23 20    90
3       23 20    56
.
.
298   108 26    89
.

> cl = cluster(algebra, c="class", 10, method="srswor")
> getdata(algebra,cl)
    Mi score class ID_unit      Prob
7   20    62    23       7 0.8333333
1   20    57    23       1 0.8333333
2   20    90    23       2 0.8333333

>attach(getdata(algebra,cl))
```

```
> yi <- tapply(score,class,sum)
> yi
  23   38   39   41   44   46   51   58   62  108
1230 1402 1972 1508 1816 1048 2308  989 1398 1746

> sum(yi)
[1] 15417
```

→ how many are in each class.

```
> mi <- tapply(score,class,length)
> mi
 23  38  39  41  44  46  51  58  62 108
 20  24  34  26  28  19  32  17  21  26

> sum(mi)
[1] 247


> ybari <- yi/mi
> ybari
      23       38       39       41       44       46       51       58       62      108
61.50000 58.41667 58.00000 58.00000 64.85714 55.15789 72.12500 58.17647 66.57143 67.15385

> ybar_r = (sum(mi*ybari)) / sum(Mi)
> ybar_r
[1] 62.417
```

→ ratio est.

```
> sum ((yi-ybari*mi)^2)
[1] 193470.1
```