Cluster Sampling: Unequal Sized

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

2/12/25

One-stage cluster sampling with unequal sizes: Implementation

- The sizes of clusters (the total number of population elements within a cluster) can vary widely in practice

One-stage cluster sampling with unequal sizes: Implementation

- The sizes of clusters (the total number of population elements within a cluster) can vary widely in practice
- We often do not know all cluster sizes in the population

One-stage cluster sampling with unequal sizes: Implementation

- The sizes of clusters (the total number of population elements within a cluster) can vary widely in practice
- We often do not know all cluster sizes in the population
- We sample n of N PSUs and take all elements within selected PSUs

One-stage cluster sampling with unequal sizes: Implementation

- The sizes of clusters (the total number of population elements within a cluster) can vary widely in practice
- We often do not know all cluster sizes in the population
- We sample n of N PSUs and take all elements within selected PSUs
- The resulting sample size is a random variable, which varies across samples

R code: Example

```r
library(sampling)
data(swissmunicipalities)
# the variable'REG' is used as clustering variable
# the sample size is 3; the method is simple random sampling without replacement
cl=cluster(swissmunicipalities,clustername=c("REG"),size=3,method="srswor")
# extracts the observed data
# the order of the columns is different from the order in the initial database
#getdata(swissmunicipalities, cl)
```

One-stage cluster sampling with unequal sizes: Inference

- We usually expect the PSU population total t_i to be positively correlated with the PSU size M_i.

One-stage cluster sampling with unequal sizes: Inference

- We usually expect the PSU population total t_i to be positively correlated with the PSU size M_i.
  - If PSUs are counties, we would expect the total number of households living in poverty in County i (t_i) to be roughly proportional to the total number of households in County i (M_i)

One-stage cluster sampling with unequal sizes: Inference

- We usually expect the PSU population total t_i to be positively correlated with the PSU size M_i.
  - If PSUs are counties, we would expect the total number of households living in poverty in County i (t_i) to be roughly proportional to the total number of households in County i (M_i)
- The population mean $\bar{Y}$ is a ratio: $\bar{Y} = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{t}{M_0} = B$

One-stage cluster sampling with unequal sizes: Inference

- We usually expect the PSU population total t_i to be positively correlated with the PSU size M_i.
  - If PSUs are counties, we would expect the total number of households living in poverty in County i (t_i) to be roughly proportional to the total number of households in County i (M_i)
- The population mean $\bar{Y}$ is a ratio: $\bar{Y} = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{t}{M_0} = B$
- The sample mean $\hat{\bar{y}} = \frac{\hat{t}}{\hat{M}_0} = \frac{\frac{N}{n}\sum_{i \in \mathcal{S}} t_i}{\frac{N}{n}\sum_{i \in \mathcal{S}} M_i}$

One-stage cluster sampling with unequal sizes: Inference

- Sampling weights $w_{ij} = \frac{1}{P(\text{SSU} j \text{ in PSU } i \text{ is selected})} = \frac{N}{n}$

One-stage cluster sampling with unequal sizes: Inference

- Sampling weights $w_{ij} = \frac{1}{P(\text{SSU}j \text{ in PSU } i \text{ is selected})} = \frac{N}{n}$
- Use the sum of weights to estimate the population size $\hat{M}_0 = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i$

One-stage cluster sampling with unequal sizes: Inference

- Sampling weights $w_{ij} = \frac{1}{P(\text{SSU}j \text{ in PSU } i \text{ is selected})} = \frac{N}{n}$
- Use the sum of weights to estimate the population size $\hat{M}_0 = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij} = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i$
- Ratio estimation for the population mean is biased $\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M_i \bar{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}$

One-stage cluster sampling with unequal sizes: Inference

- Sampling weights $w_{ij} = \frac{1}{P(\text{SSU}j \text{ in PSU } i \text{ is selected})} = \frac{N}{n}$
- Use the sum of weights to estimate the population size $\hat{M}_0 = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i$
- Ratio estimation for the population mean is biased $\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}$
- Taking $e_i = t_i - M_i \hat{\bar{y}}_r = M_i(\bar{y}_i - \hat{\bar{y}}_r)$ based on the MSE of the ratio mean, we have $s_r^2 = \frac{1}{n-1} \sum_i M_i^2 (\bar{y}_i - \hat{\bar{y}}_r)^2$ and $SE(\hat{\bar{y}}_r) = \sqrt{(1 - \frac{n}{N}) \frac{s_r^2}{n \bar{M}^2}}$

Two-stage cluster sampling with unequal sizes: Implementation

1. Select an SRS of n PSUs from the population of N PSUs, with the first stage selection rate $f_n = \frac{n}{N}$

Two-stage cluster sampling with unequal sizes: Implementation

1. Select an SRS of n PSUs from the population of N PSUs, with the first stage selection rate $f_n = \frac{n}{N}$
2. Selection an SRS of m_i SSUs from the selected PSU i, with the second stage selection rate $f_m = \frac{m_i}{M_i}$

Two-stage cluster sampling with unequal sizes: Implementation

1. Select an SRS of n PSUs from the population of N PSUs, with the first stage selection rate $f_n = \frac{n}{N}$
2. Selection an SRS of m_i SSUs from the selected PSU i, with the second stage selection rate $f_m = \frac{m_i}{M_i}$

- Overall two-stage sampling rate $f = f_n * f_m = \frac{nm_i}{NM_i}$

Two-stage cluster sampling with unequal sizes: Implementation

1. Select an SRS of n PSUs from the population of N PSUs, with the first stage selection rate $f_n = \frac{n}{N}$
2. Selection an SRS of m_i SSUs from the selected PSU i, with the second stage selection rate $f_m = \frac{m_i}{M_i}$

- Overall two-stage sampling rate $f = f_n * f_m = \frac{nm_i}{NM_i}$
- Sampling weights $w_{ij} = \frac{NM_i}{nm_i}$

Example

- Fixed two-stage sampling rates $f_n = \frac{1}{2}$ and $f_m = \frac{1}{10}$ (i.e., $f = \frac{1}{20}$) from a population of $N = 10$ unequal sized clusters with $M_0 = 1850$, with an average sample size of $f * M_0 = 92.5$

(A) Two unequal cluster samples, $f = 1/20$

Cluster	Size	Sample 1			Sample 2		
α	B_α	α	B_α	$x_\alpha = f_b B_\alpha$	α	B_α	$x_\alpha = f_b B_\alpha$
1	400	1	400	40			
2	310				2	310	31
3	40	3	40	4			
4	150				4	150	15
5	250	5	250	25			
6	220				6	220	22
7	50	7	50	5			
8	130				8	130	13
9	90	9	90	9			
10	210				10	210	21
Total	1850	5	830	83	5	1020	102

Variation in sample size

- With unequal-sized clusters and fixed sampling rates at both stages of selection, our achieved sample size will randomly vary across hypothetical samples (despite the epsem selection)! Our **sample size is a random variable**

Variation in sample size

- With unequal-sized clusters and fixed sampling rates at both stages of selection, our achieved sample size will randomly vary across hypothetical samples (despite the epsem selection)! Our **sample size is a random variable**
- If we were to treat the achieved sample size as fixed, we would be failing to recognize the variation in the sample size when estimating variance, and we would underestimate the sampling variance

Variation in sample size

- With unequal-sized clusters and fixed sampling rates at both stages of selection, our achieved sample size will randomly vary across hypothetical samples (despite the epsem selection)! Our **sample size is a random variable**
- If we were to treat the achieved sample size as fixed, we would be failing to recognize the variation in the sample size when estimating variance, and we would underestimate the sampling variance
- We could depart from epsem to eliminate variation in sample size, selecting a fixed subsample size from each cluster (need weights)

Variation in sample size

- With unequal-sized clusters and fixed sampling rates at both stages of selection, our achieved sample size will randomly vary across hypothetical samples (despite the epsem selection)! Our **sample size is a random variable**
- If we were to treat the achieved sample size as fixed, we would be failing to recognize the variation in the sample size when estimating variance, and we would underestimate the sampling variance
- We could depart from epsem to eliminate variation in sample size, selecting a fixed subsample size from each cluster (need weights)
- Better solution: Probability Proportionate to Size (PPS) sampling (discuss later)

Two-stage cluster sampling with unequal sizes: Inference

- The population total estimator $\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$

Two-stage cluster sampling with unequal sizes: Inference

- The population total estimator $\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$
- The sampling variance
  $Var(\hat{t}) = N^2 (1 - \frac{n}{N}) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} (1 - \frac{m_i}{M_i}) M_i^2 \frac{S_i^2}{m_i}$

Two-stage cluster sampling with unequal sizes: Inference

- The population total estimator $\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$
- The sampling variance
  $Var(\hat{t}) = N^2 (1 - \frac{n}{N}) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} (1 - \frac{m_i}{M_i}) M_i^2 \frac{S_i^2}{m_i}$
- The population mean estimation $\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}$

Two-stage cluster sampling with unequal sizes: Inference

- The population total estimator $\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$
- The sampling variance
  $Var(\hat{t}) = N^2(1 - \frac{n}{N})\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}(1 - \frac{m_i}{M_i})M_i^2\frac{S_i^2}{m_i}$
- The population mean estimation $\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}$
- The sampling variance estimator for the ratio mean
  $var(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}(1 - \frac{n}{N})\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i \in \mathcal{S}} M_i^2(1 - \frac{m_i}{M_i})\frac{s_i^2}{m_i}$

Simplified variance

- The second term is generally small compared to the first term. The MSE and variance are equal. Use Taylor Series Linearization.

Simplified variance

- The second term is generally small compared to the first term. The MSE and variance are equal. Use Taylor Series Linearization.
- The first stage selection is assumed as with replacement.

Simplified variance

- The second term is generally small compared to the first term. The MSE and variance are equal. Use Taylor Series Linearization.
- The first stage selection is assumed as with replacement.
- Most survey software packages use this as the sampling variable for $\hat{\bar{y}}_r = \frac{\hat{t}}{\hat{M}_0} = r = \frac{\hat{t}_y}{\hat{t}_x}$

$$
\begin{aligned}
var(\hat{\bar{y}}_r) &\approx \frac{s_r^2}{n\bar{M}^2} = \frac{1}{(n-1)n\bar{M}^2}\sum_i M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2 \\
&\approx \frac{1}{\hat{t}_x^2}[var(\hat{t}_y) + r^2 var(\hat{t}_x) - 2 * r * cov(\hat{t}_y, \hat{t}_x)] \\
&= \frac{1}{\hat{t}_x^2}\frac{n(1-f)}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2\sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i})
\end{aligned}
$$

where $\hat{t}_{y,i}$ and $\hat{t}_{x,i}$ is the weighted cluster total of the measure values and sample size, respectively.

Adequacy of approximation

- Since $var(\hat{\bar{y}}_r)$ is an approximation, it would be useful to an indication of when it might fail

Adequacy of approximation

- Since $var(\hat{\bar{y}}_r)$ is an approximation, it would be useful to an indication of when it might fail
- Examination of the Taylor series shows the adequacy of the approximation depends on the coefficient of variation of the sample size (i.e., sum of weights within clusters):

$$cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = \frac{\sqrt{\frac{n(1-f)}{n-1}[\sum \hat{t}_{x,i}^2 - \frac{(\sum \hat{t}_{x,i})^2}{n}]}}{\sum_i \hat{t}_{x,i}}$$

Adequacy of approximation

- Since $var(\hat{\bar{y}}_r)$ is an approximation, it would be useful to an indication of when it might fail

- Examination of the Taylor series shows the adequacy of the approximation depends on the coefficient of variation of the sample size (i.e., sum of weights within clusters):

$$cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = \frac{\sqrt{\frac{n(1-f)}{n-1}[\sum \hat{t}_{x,i}^2 - \frac{(\sum \hat{t}_{x,i})^2}{n}]}}{\sum_i \hat{t}_{x,i}}$$

- As long as $cv(\hat{t}_x) < 0.1$, the approximation is reasonably accurate: Values of $cv(\hat{t}_x)$ as large as 0.2 may also be acceptable

Adequacy of approximation

- Since $var(\hat{\bar{y}}_r)$ is an approximation, it would be useful to an indication of when it might fail

- Examination of the Taylor series shows the adequacy of the approximation depends on the coefficient of variation of the sample size (i.e., sum of weights within clusters):

$$cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = \frac{\sqrt{\frac{n(1-f)}{n-1}[\sum \hat{t}^2_{x,i} - \frac{(\sum \hat{t}_{x,i})^2}{n}]}}{\sum_i \hat{t}_{x,i}}$$

- As long as $cv(\hat{t}_x) < 0.1$, the approximation is reasonably accurate: Values of $cv(\hat{t}_x)$ as large as 0.2 may also be acceptable

- The bias of the ratio estimator r also depends on the coefficient of variation of the denominator: $\mid \frac{Bias(r)}{se(r)} \mid < cv(\hat{t}_x)$

Unequal-sized clusters: Projection

- Based on the variance estimate and the SRS variance estimate (with the same sample size), we can estimate DEFF

Unequal-sized clusters: Projection

- Based on the variance estimate and the SRS variance estimate (with the same sample size), we can estimate DEFF
- Then we use the average cluster size ($\bar{m}$) and can estimate roh, for thinking about new sample designs

Unequal-sized clusters: Projection

- Based on the variance estimate and the SRS variance estimate (with the same sample size), we can estimate DEFF
- Then we use the average cluster size ($\bar{m}$) and can estimate roh, for thinking about new sample designs
- We can compute a new design effect ($DEFF = 1 + (m–1) * roh$) given new choices of n and m

Unequal-sized clusters: Projection

- Based on the variance estimate and the SRS variance estimate (with the same sample size), we can estimate DEFF
- Then we use the average cluster size ($\bar{m}$) and can estimate roh, for thinking about new sample designs
- We can compute a new design effect ($DEFF = 1 + (m–1) * roh$) given new choices of n and m
- We then multiply the NEW SRS variance (given portable estimates and total sample size $n * m$) by the NEW DEFF to get the NEW expected sampling variance

Unequal-sized clusters: Projection

- Based on the variance estimate and the SRS variance estimate (with the same sample size), we can estimate DEFF
- Then we use the average cluster size ($\bar{m}$) and can estimate roh, for thinking about new sample designs
- We can compute a new design effect ($DEFF = 1 + (m–1) * roh$) given new choices of n and m
- We then multiply the NEW SRS variance (given portable estimates and total sample size $n * m$) by the NEW DEFF to get the NEW expected sampling variance
- This is no different from what we've done before!

Example

- In a hospital authority of 5 hospitals, estimate the proportion of outpatient visits due to trauma
  - Element: outpatient visits
  - Estimate: proportion of visits due to trauma
- Select outpatient records in 2 stages:
  - First hospital-days ($N = 5 * 365 = 1825$)
    - Clusters: days across hospitals
    - Sample of $n = 10$ hospital-days
  - Second, select all records on a selected day

$$f = \frac{10}{1825} * 1$$

Hospital-days

Sample hospital-day α	Total visits x_α	Trauma visits y_α
1	58	40
2	47	16
3	37	8
4	69	27
5	40	10
6	27	18
7	34	17
8	30	12
9	26	16
10	32	16

$\sum x_\alpha = 400$ $\sum y_\alpha = 180$ $\sum x_\alpha^2 = 17778$ $\sum y_\alpha^2 = 4018$ $\sum x_\alpha y_\alpha = 7983$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2 \sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2 \sum_i \hat{t}_{x,i}^2 - 2r \sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2} \frac{n}{n-1} (\sum_i \hat{t}_{y,i}^2 + r^2 \sum_i \hat{t}_{x,i}^2 - 2r \sum_i \hat{t}_{y,i} \hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$
- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1114$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2 \sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$
- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1114$
- deff: $var_{srs}(r) = \frac{r(1-r)}{\hat{t}_x - 1} = 0.0006202$, so $deff = \frac{0.003023}{0.0006202} = 4.874$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2\sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$
- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1114$
- deff: $var_{srs}(r) = \frac{r(1-r)}{\hat{t}_x - 1} = 0.0006202$, so $deff = \frac{0.003023}{0.0006202} = 4.874$
- $roh = \frac{deff-1}{\bar{m}-1} = \frac{4.874-1}{400/10-1} = 0.09934$

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2 \sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$
- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1114$
- deff: $var_{srs}(r) = \frac{r(1-r)}{\hat{t}_x - 1} = 0.0006202$, so $deff = \frac{0.003023}{0.0006202} = 4.874$
- $roh = \frac{deff-1}{\bar{m}-1} = \frac{4.874-1}{400/10-1} = 0.09934$
- What about a new sample design with $n = 20$ and $m = 20$?

Example: cont.

- The ratio mean $r = 0.45$ with $var(r) = \frac{1}{\hat{t}_x^2}\frac{n}{n-1}(\sum_i \hat{t}_{y,i}^2 + r^2\sum_i \hat{t}_{x,i}^2 - 2r\sum_i \hat{t}_{y,i}\hat{t}_{x,i}) = 0.003023$
- With $t_{0.975,9} = 2.262$, the 95% CI is $(0.3256, 0.5744)$
- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1114$
- deff: $var_{srs}(r) = \frac{r(1-r)}{\hat{t}_x - 1} = 0.0006202$, so $deff = \frac{0.003023}{0.0006202} = 4.874$
- $roh = \frac{deff-1}{\bar{m}-1} = \frac{4.874-1}{400/10-1} = 0.09934$
- What about a new sample design with $n = 20$ and $m = 20$?
    - Compute new deff, and multiply by new SRS sampling variance to obtain new sampling variance

Example: Lohr 5.7

```r
### With-replacement variance
# calculate with-replacement variance; no fpc argument
# include psu variable in id; include weights
dschools<-svydesign(id=~schoolid,weights=~finalwt,data=schools); dschools
```

```
1 - level Cluster Sampling design (with replacement)
With (10) clusters.
svydesign(id = ~schoolid, weights = ~finalwt, data = schools)
```

```r
# dschools tells you this is treated as a with-replacement sample
mathmean<-svymean(~math,dschools); mathmean
```

```
       mean     SE
math 33.123 1.7599
```

```r
degf(dschools)
```

```
[1] 9
```

```r
# use t distribution for confidence intervals because there are only 10 psus
confint(mathmean,df=degf(dschools))
```

```
        2.5 %  97.5 %
math 29.14179 37.1041
```

Example: Lohr 5.7 cont.

```
### Without-replacement variance
# create a variable giving each student an id number
schools$studentid<-1:(nrow(schools))
# specify both stages of the sample in the id argument
# give both sets of population sizes in the fpc argument
# do not include the weight argument
dschoolwor<-svydesign(id=~schoolid+studentid,fpc=~rep(75,nrow(schools))+Mi,
                       data=schools)
dschoolwor
```

```
2 - level Cluster Sampling design
With (10, 200) clusters.
svydesign(id = ~schoolid + studentid, fpc = ~rep(75, nrow(schools)) +
    Mi, data = schools)
```

```
mathmeanwor<-svymean(~math,dschoolwor); mathmeanwor
```

```
       mean     SE
math 33.123 1.6605
```

```
confint(mathmeanwor,df=degf(dschoolwor))
```

```
        2.5 %   97.5 %
math 29.36667 36.87923
```

Example: Lohr 5.7 cont.

- The adequacy of the approximation: $cv(\hat{t}_x) = \frac{se(\hat{t}_x)}{\hat{t}_x} = 0.1212787$

```
### cv of sample sizes
Mi = unique(schools$Mi); Mi
```

```
 [1] 163 180 114 367 109 219 318 259 311 263
```

```
n = length(Mi)
se_Mi = sqrt(var(Mi) * n) #se of the total
se_Mi/sum(Mi) #cv
```

```
[1] 0.1212787
```

```
sqrt(n/(n-1) * (sum(Mi^2) - sum(Mi)^2/n))/sum(Mi) # use formula
```

```
[1] 0.1212787
```

Stratified cluster sampling

- We apply the same basic stratified sampling technique to select samples of unequal-sized clusters from within strata
- Cluster sample stratification similar to elements
  - Use cluster characteristics to stratify clusters
  - Homogeneous, mutually exclusive, exhaustive
  - Control the distribution of the sample
  - Decrease sampling variance
  - Stratifying variables, boundaries, etc., discussed for element sampling applies to clusters

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!
- All of the same stratification concepts that we discussed for elements applies in the same way to clusters (unequal sizes or not)

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!
- All of the same stratification concepts that we discussed for elements applies in the same way to clusters (unequal sizes or not)
- Allocation of sample clusters across strata:

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!
- All of the same stratification concepts that we discussed for elements applies in the same way to clusters (unequal sizes or not)
- Allocation of sample clusters across strata:
    - Proportionate allocation: usually refers to elements and not clusters; This kind of allocation allows us to maintain epsem for the elements (not the clusters themselves)

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!
- All of the same stratification concepts that we discussed for elements applies in the same way to clusters (unequal sizes or not)
- Allocation of sample clusters across strata:
  - Proportionate allocation: usually refers to elements and not clusters; This kind of allocation allows us to maintain epsem for the elements (not the clusters themselves)
  - Paired selection: facilitates variance estimation, and as many strata as possible, but adds constraints to the design

Stratified cluster sampling: Implementation

- We use cluster characteristics to stratify clusters; if these are highly correlated with individual characteristics, that is optimal!
- All of the same stratification concepts that we discussed for elements applies in the same way to clusters (unequal sizes or not)
- Allocation of sample clusters across strata:
  - Proportionate allocation: usually refers to elements and not clusters; This kind of allocation allows us to maintain epsem for the elements (not the clusters themselves)
  - Paired selection: facilitates variance estimation, and as many strata as possible, but adds constraints to the design
  - Other allocations

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$
- Variance estimation depends on the methods used to select the clusters **within each stratum**

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$
- Variance estimation depends on the methods used to select the clusters **within each stratum**
  - Based on ultimate cluster sampling theory: only first-stage components will be used for variance estimation

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$
- Variance estimation depends on the methods used to select the clusters **within each stratum**
  - Based on ultimate cluster sampling theory: only first-stage components will be used for variance estimation
  - Under disproportionate allocation, $(1 - f_h)$ appears in each stratum

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$
- Variance estimation depends on the methods used to select the clusters **within each stratum**
  - Based on ultimate cluster sampling theory: only first-stage components will be used for variance estimation
  - Under disproportionate allocation, $(1 - f_h)$ appears in each stratum
  - Multiple, paired, successive differences

Stratified cluster sampling: Inference

- Add subscripts: Stratum h, PSU i, SSU j
- Ratio mean $\bar{y} = \frac{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij} y_{hij}}{\sum_h \sum_{i \in h} \sum_{j \in \mathcal{S}_i} w_{hij}}$
- Variance estimation depends on the methods used to select the clusters **within each stratum**
    - Based on ultimate cluster sampling theory: only first-stage components will be used for variance estimation
    - Under disproportionate allocation, $(1 - f_h)$ appears in each stratum
    - Multiple, paired, successive differences
- Perfectly fine to have different variance contributions from different strata, depending on the type of sampling conducted

Multiple differences

- The same sampling variance for the ratio estimator under TSL

$$
\begin{aligned}
var(r) &\approx \frac{1}{\hat{t}_x^2}[var(\hat{t}_y) + r^2 var(\hat{t}_x) - 2 * r * cov(\hat{t}_y, \hat{t}_x)] \\
&= \frac{1}{\hat{t}_x^2}[\sum_h var(\hat{t}_{h,y}) + r^2 \sum_h var(\hat{t}_{h,x}) - 2r \sum_h cov(\hat{t}_{h,y}, \hat{t}_{h,x})]
\end{aligned}
$$

Paired selection

- For all strata $n_h = 2$

$$
\begin{aligned}
var(r) \approx & \frac{1}{\hat{t}_x^2}[\sum_h var(\hat{t}_{h,y}) + r^2 \sum_h var(\hat{t}_{h,x}) - 2r \sum_h cov(\hat{t}_{h,y}, \hat{t}_{h,x})] \\
= & \frac{1}{\hat{t}_x^2}[\sum_h (1-f_h)(\hat{t}_{h,1,y} - \hat{t}_{h,2,y})^2 + \\
& r^2 \sum_h (1-f_h)(\hat{t}_{h,1,x} - \hat{t}_{h,2,x})^2 - \\
& 2r \sum_h (1-f_h)(\hat{t}_{h,1,y} - \hat{t}_{h,2,y})(\hat{t}_{h,1,x} - \hat{t}_{h,2,x})]
\end{aligned}
$$

- Actual selection may be: Paired selection design; Systematic selection collapsed to paired differences (discussed later); One selection per stratum collapsed to pairs

Successive differences

- Successive differences ($n_h > 2$ for all strata):

$$
\begin{aligned}
var(r) \approx & \frac{1}{\hat{t}_x^2}[\sum_h \frac{n_h(1-f_h)}{2(n_h-1)} \sum_{g=1}^{n_h-1} (\hat{t}_{h,g,y} - \hat{t}_{h,g+1,y})^2 + \\
& r^2 \sum_h \frac{n_h(1-f_h)}{2(n_h-1)} \sum_{g=1}^{n_h-1} (\hat{t}_{h,g,x} - \hat{t}_{h,g+1,x})^2 - \\
& 2r \sum_h \frac{n_h(1-f_h)}{2(n_h-1)} \sum_{g=1}^{n_h-1} (\hat{t}_{h,g,y} - \hat{t}_{h,g+1,y})(\hat{t}_{h,g,x} - \hat{t}_{h,g+1,x})]
\end{aligned}
$$

- Actual selection may be systematic selection from ordered list

Using ratio estimation results

- Taking $e_i = t_i - M_i\hat{\bar{y}}_r = M_i(\bar{y}_i - \hat{\bar{y}}_r)$ based on the MSE of the ratio mean, we have $s_r^2 = \frac{1}{n-1}\sum_i M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2$ and $var(\hat{\bar{y}}_r) = \frac{s_r^2}{n\bar{M}^2}$
- Let $e_{hi} = y_{hi} - rx_{hi}$, we have alternative formulations of these three formulas

$$var(r) \approx \frac{1}{x^2}[\sum_h \frac{n_h(1-f_h)}{n_h-1}(\sum_{i=1}^{n_h} e_{hi}^2 - \frac{(\sum_{i=1}^{n_h} e_{hi})^2}{n_h})]$$

$$var(r) \approx \frac{1}{x^2}[\sum_h (1-f_h)(e_{h1} - e_{h2})^2]$$

$$var(r) \approx \frac{1}{x^2}[\sum_h \frac{n_h(1-f_h)}{2(n_h-1)} \sum_{g=1}^{n_h-1} (e_{hg} - e_{h,g+1})^2]$$

Example: Paired selection

h (Stratum)	α (SECU)	$y_{h\alpha}$	$x_{h\alpha}$	f_h
31	1	299	41	0.47
31	2	680	100	
42	1	67	7	0.24
42	2	49	5	
35	1	125	33	0.09
35	2	64	14	

- SECU: sampling error computation units, in a similar role with clusters or primary sampling units, will be discussed in detail later

Example: cont.

$$r = \frac{\sum_h \sum_\alpha y_{h\alpha}}{\sum_h \sum_\alpha x_{h\alpha}} = \frac{1284}{200} = 6.42$$

$$\begin{aligned}
\operatorname{var}(r) &\approx \frac{1}{x^2}\left[\operatorname{var}(y) + r^2 \operatorname{var}(x) - 2r\operatorname{cov}(y,x)\right] \\
&= \frac{1}{x^2}\left[\begin{array}{l} \sum_h (1-f_h)(y_{h1}-y_{h2})^2 + r^2 \sum_h (1-f_h)(x_{h1}-x_{h2})^2 \\ -2r\sum_h (1-f_h)(y_{h1}-y_{h2})(x_{h1}-x_{h2}) \end{array}\right] \\
&= \frac{1}{200^2}\left[80567.68 + 6.42^2 \times 2176.48 - 2(6.42)(12995.92)\right] \\
&= 0.0852
\end{aligned}$$

$$se(r) = \sqrt{0.0852} = 0.2918$$

$$df = a - H = 6 - 3 = 3$$

$$t_{0.975,3} = 3.18$$

95% CI L.L. = $6.42 - 3.18 \times 0.2918 = 5.4920$

95% CI U.L. = $6.42 + 3.18 \times 0.2918 = 7.3480$

$$cv(x) = \frac{se(x)}{x} = \frac{46.6528}{200} = 0.2333$$

Summary

- With unequal-sized clusters and fixed sampling rates at both stages of selection, our achieved sample size will randomly vary across hypothetical samples (despite the epsem selection)! Our sample size is a random variable.
- Two main problems with ratio means:
  - They are biased estimators of the overall population mean!
  - Theoretical sampling variance of the ratio mean is not known exactly!
- Remember that the estimated variance of the ratio mean is an approximation
  - Key diagnostic quantity: cv of the achieved sample size
- Stratified unequal-sized cluster sampling
  - Independent two-stage sampling (cluster and elements) across strata