

Cluster Sampling: Equal Sized

SurvMeth/Surv 625: Applied Sampling

Yajuan Si

University of Michigan, Ann Arbor

2/5/25

Poststratification: Example revisit

- An SRS of 400 students taken from a school with 4000 students: 240 women and 160 men
- With 84 of the sampled women and 40 of the sampled men planning to follow careers in academia
- Question: How many students planning to work in academia?

SRS inference

- Population total estimate:

$$\hat{t} = N * p = 4000 * \frac{124}{400} = \frac{4000}{400} * 124 = 1240$$

- The sampling variance of \hat{t} :

$$\begin{aligned} \text{var}(\hat{t}) &= N^2 * \text{var}(p) = N^2 * \left(1 - \frac{n}{N}\right) * \frac{s^2}{n} & (1) \\ &= N^2 * \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1} \\ &\approx N^2 * \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n} \\ &= 4000^2 * \left(1 - \frac{400}{4000}\right) \frac{\frac{124}{400} \left(1 - \frac{124}{400}\right)}{400} \end{aligned}$$

Poststratification (Holt and Smith 1979)

- 1 If we know that the school has 2700 women and 1300 men, we construct two poststrata and estimate within each stratum, where $p_1 = \frac{84}{240}$, $p_2 = \frac{40}{160}$, $s_1^2 = \frac{n_1}{n_1-1}p_1(1-p_1)$, and $s_2^2 = \frac{n_2}{n_2-1}p_2(1-p_2)$
- 2 We combine the two stratum-wide estimates to obtain the poststratification estimator

$$\bar{t}_{post} = \sum_h N_h p_h = 2700 * \frac{84}{240} + 1300 * \frac{40}{160} = 1270$$

- 3 Because the sample size within stratum n_h is random, $E(\frac{1}{n_h}) \approx \frac{1}{nW_h} + \frac{1-W_h}{n^2W_h^2}$, we need to account for its variability in the sampling variance estimation

$$var(\bar{t}_{post}) = \frac{1-f}{n} \sum_h N^2 W_h s_h^2 + \frac{1}{n^2} \sum_h N^2 (1-W_h) s_h^2. \quad (2)$$

The second term is usually small if n/H is large.

Stratified sampling

- 1 If we have had implemented stratified sampling: We independently select women and men with a SRS of 240 women out of 2700 and another SRS of 160 men out of 1300
- 2 The stratified estimate is

$$\bar{t}_{st} = \sum_h N_h p_h = 2700 * \frac{84}{240} + 1300 * \frac{40}{160} = 1270$$

- 3 The sampling variance

$$var(\bar{t}_{st}) = \sum_h N^2 W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$













The sampling variance has to accurately reflect the sampling process!

Cluster sampling: Implementation

- Populations often distributed geographically
 - Cannot afford to create an element frame
 - Cannot afford to visit all units drawn randomly from the entire area
- Save costs: SRS tends to be prohibitively expensive relative to cluster sampling
- Clusters are naturally occurring, e.g., for human populations, clustering factors include
 - Environment (exposure to infectious disease)
 - Self-selection (poor households in same block)
 - Interaction (shared attitudes among neighbors)
- Many frames only list clusters and elements only for selected clusters
- Clusters are the primary sampling units (PSUs), and the secondary sampling units (SSUs) are the elements

Comparison between cluster and stratified sampling

- Clusters are randomly selected, while strata are not sampled
- Clusters are naturally occurring units, while strata are fixed by design based on auxiliary information
- One stratum may include elements from multiple clusters
- The number of clusters is often large, while the number of strata is often between 3 and 6.

Clusters/ Strata:	Clustering	Stratification
A		
B		
C		
D		
E		
F		

Comparison: cont.

	Stratified sampling	Cluster sampling
Meaning	The researcher divides population into a few strata, and then samples are randomly selected within each stratum	Naturally occurring groups called 'clusters' are identified (possibly on a frame), and then clusters are sampled
Sample	Individuals selected from all strata	Collectively from randomly selected clusters
In order to minimize sampling error	Within-group differences among strata should be minimized, and between-group differences among strata should be maximized	Within-group differences should be consistent with those in the population, and between-group differences among the clusters should be minimized (will minimize roh!)
Sampling frame	A sampling frame is needed for the entire target population	In single-stage cluster sampling, a sampling frame is needed only for the clusters. In two-stage and multistage cluster sampling, a sampling frame of individual elements is needed only for the elements in selected clusters

Comparison: cont.

	Stratified sampling	Cluster sampling
Objective	Increase precision and representation	Decrease costs and increase operational efficiency
Bifurcation	Imposed by the researcher	Naturally occurring groups
Compared with SRS	More precision (generally)	Less precision
Correlation between grouping variables and outcome variable	Should be related to the research problem	Should not be related to the research problem
Common grouping variables	Age, gender, income, race (aggregated at geographies)	Geographical area, school, classroom, institution

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

SSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

SSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

SSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

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

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

SSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

$$t = \sum_{i=1}^N t_i =$$

$$\sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$$

- Population variance of the PSU totals: $S_t^2 =$

$$\frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

$$t = \sum_{i=1}^N t_i =$$

$$\sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$$

- Population variance of the PSU totals: $S_t^2 =$

$$\frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$$

SSU level

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

$$t = \sum_{i=1}^N t_i =$$

$$\sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$$

- Population variance of the PSU totals: $S_t^2 =$

$$\frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$$

SSU level

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

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

- Population variance of the PSU totals: $S_t^2 = \frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$

SSU level

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y}_i)^2}{M_i - 1}$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

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

- Population variance of the PSU totals: $S_t^2 = \frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$

SSU level

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y}_i)^2}{M_i - 1}$$

- Average within-PSU variance:

$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$

Notation: Population quantities

The population has N PSUs, the number of SSUs in the i th PSU is M_i , and the measurement for the j th element in i th PSU is y_{ij} . Population

mean: $\bar{Y} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$.

PSU level

- Total # SSUs:

$$M_0 = \sum_{i=1}^N M_i$$

- Total in PSU i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Population total:

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

- Population variance of the PSU totals: $S_t^2 = \frac{1}{N-1} \sum_{i=1}^N (t_i - t/N)^2$

SSU level

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{Y}_i)^2}{M_i - 1}$$

- Average within-PSU variance:

$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$

- Between PSU variance:

$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :

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

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :
$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$
- Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i y_{ij}}{m_i}$$

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :
$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$
- Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i y_{ij}}{m_i}$$
- Estimated population total:
$$\hat{t} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i$$

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :
$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$
- Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i y_{ij}}{m_i}$$
- Estimated population total:
$$\hat{t} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i$$
- $s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - \hat{t}/n)^2$

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :
$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$
- Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i y_{ij}}{m_i}$$
- Estimated population total:
$$\hat{t} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i$$
- $s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - \hat{t}/n)^2$
- Sample element variance within PSU i : $s_i^2 = \sum_{j \in \mathcal{S}_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1}$

Notation: Sample quantities

- The sample \mathcal{S} has n PSUs
- m_i is the number of SSUs selected from the i th PSU
- \mathcal{S}_i is the sample of selected SSUs within PSU i
- Sample mean for PSU i :
$$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$$
- Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i y_{ij}}{m_i}$$
- Estimated population total:
$$\hat{t} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i$$
- $s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - \hat{t}/n)^2$
- Sample element variance within PSU i : $s_i^2 = \sum_{j \in \mathcal{S}_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1}$
- Sampling weight for SSU j in PSU i : w_{ij}

Inference: One-stage cluster sampling with equal sizes

- Consider the simplest case when each PSU has the same number of elements $M_i = m_i = M$ and we take all elements within selected PSUs

Inference: One-stage cluster sampling with equal sizes

- Consider the simplest case when each PSU has the same number of elements $M_i = m_i = M$ and we take all elements within selected PSUs
- We have an SRS of n data points, $\{t_i, i \in S\}$, with t_i being the total in PSU i

Inference: One-stage cluster sampling with equal sizes

- Consider the simplest case when each PSU has the same number of elements $M_i = m_i = M$ and we take all elements within selected PSUs
- We have an SRS of n data points, $\{t_i, i \in S\}$, with t_i being the total in PSU i
- The population mean estimate $\hat{y} = \frac{\sum_i t_i}{nM}$ with
 $SE(\hat{y}) = \frac{1}{M} \sqrt{(1 - \frac{n}{N}) \frac{s_t^2}{n}}$ where $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (t_i - \hat{t}/N)^2$

Inference: One-stage cluster sampling with equal sizes

- Consider the simplest case when each PSU has the same number of elements $M_i = m_i = M$ and we take all elements within selected PSUs
- We have an SRS of n data points, $\{t_i, i \in S\}$, with t_i being the total in PSU i
- The population mean estimate $\hat{y} = \frac{\sum_i t_i}{nM}$ with
 $SE(\hat{y}) = \frac{1}{M} \sqrt{(1 - \frac{n}{N}) \frac{s_t^2}{n}}$ where $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (t_i - \hat{t}/N)^2$
- The sampling variance **only involves between-cluster variance**, similar to the SRS sampling variance.

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites
- The inference only needs the GPA totals across 5 PSUs are (12.16, 11.36, 8.96, 12.96, 11.08)

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites
- The inference only needs the GPA totals across 5 PSUs are (12.16, 11.36, 8.96, 12.96, 11.08)
- The estimate of the population total
$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + \cdots + 11.08) = 1130.4$$

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites
- The inference only needs the GPA totals across 5 PSUs are (12.16, 11.36, 8.96, 12.96, 11.08)
- The estimate of the population total
$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + \dots + 11.08) = 1130.4$$
- The sampling variance of totals
$$\hat{s}_t^2 = \frac{1}{5-1}[(12.16 - 11.304)^2 + \dots] = 2.256$$

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites
- The inference only needs the GPA totals across 5 PSUs are (12.16, 11.36, 8.96, 12.96, 11.08)
- The estimate of the population total
$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + \dots + 11.08) = 1130.4$$
- The sampling variance of totals
$$\hat{s}_t^2 = \frac{1}{5-1}[(12.16 - 11.304)^2 + \dots] = 2.256$$
- The mean $\hat{y} = 1130.4/400 = 2.826$ with
$$SE(\hat{y}) = \frac{1}{4}\sqrt{(1 - 5/100)2.256/5} = 0.164$$

Inference: Lohr Example 5.2

- A random selection of 5 out of 100 suites with GPA of all 4 persons in the 5 suites
- The inference only needs the GPA totals across 5 PSUs are (12.16, 11.36, 8.96, 12.96, 11.08)
- The estimate of the population total
$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + \dots + 11.08) = 1130.4$$
- The sampling variance of totals
$$\hat{s}_t^2 = \frac{1}{5-1}[(12.16 - 11.304)^2 + \dots] = 2.256$$
- The mean $\hat{y} = 1130.4/400 = 2.826$ with
$$SE(\hat{y}) = \frac{1}{4}\sqrt{(1 - 5/100)2.256/5} = 0.164$$
- A 95% CI for the mean is $2.826 \pm t_{4,0.975} * 0.164$

R code

```
# define one-stage cluster design
# note that id is suite instead of individual student as we take an SRS of suites
dgpa<-svydesign(id=~suite,weights=~wt,fpc=~rep(100,20),data=gpa); dgpa
```

1 - level Cluster Sampling design

With (5) clusters.

```
svydesign(id = ~suite, weights = ~wt, fpc = ~rep(100, 20), data = gpa)
```

estimate mean and se

```
gpamean<-svymean(~gpa,dgpa); gpamean
```

```
      mean      SE
gpa 2.826 0.1637
degf(dgpa)
```

```
[1] 4
```

n=5, t-approximation is suggested for CI

```
confint(gpamean,level=.95,df=4) # use t-approximation
```

```
      2.5 %   97.5 %
gpa 2.371593 3.280407
```

```
# confint(gpamean,level=.95) # uses normal approximation, if desired (for large n)
```

ANOVA: Cluster sampling

Source	Mean Square
Between PSUs	$MSB = \frac{SSB}{N-1} = \frac{\sum_{i=1}^N \sum_{j=1}^M (\bar{Y}_i - \bar{Y})^2}{N-1}$
Within PSU	$MSW = \frac{SSW}{N(M-1)} = \frac{\sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2}{N(M-1)}$
Total	$S^2 = \frac{SSTO}{NM-1} = \frac{\sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y})^2}{NM-1}$

- Cluster sampling only depends on the between PSUs variability:

$$V(\hat{t}_{cluster}) = N^2 \left(1 - \frac{n}{N}\right) \frac{M * MSB}{n}$$

Intraclass correlation coefficient (ICC)

- The ICC measures how similar elements in the same cluster are, a measure of homogeneity or rate of homogeneity (*roh*)
- Defined as the ratio of the between-cluster variance to the sum of the between- and within-cluster variances $ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$
- With ANOVA, we have $MSB = M\sigma_B^2 + \sigma_W^2$ and $MSW = \sigma_W^2$, so $ICC = \frac{MSB - MSW}{MSB + (M-1)MSW}$ and $-\frac{1}{M-1} \leq ICC \leq 1$
- If clusters are perfectly homogeneous, then $MSW = 0$ and $ICC = 1$
- The adjusted R_a^2 in linear regression, $R_a^2 = 1 - \frac{MSW}{S^2}$, is close to the ICC, the relative amount of population variability explained by the PSU means

Comparing to an SRS of the sample size

- $V(\hat{t}_{SRS}) = (NM)^2(1 - \frac{nM}{NM}) \frac{S^2}{nM} = N^2(1 - \frac{n}{N}) \frac{M*S^2}{n}$
- The deff, i.e., the ratio of two sampling variances, more complex cluster sampling variance in the numerator, and SRS of the same size variance in the denominator

$$deff = \frac{V(\hat{t}_{cluster})}{V(\hat{t}_{SRS})} = \frac{MSB}{S^2}$$

$$deff = 1 + (M - 1) * roh \quad (3)$$

- The deff of cluster sampling is often larger than 1

Two-stage cluster sampling with equal sizes

- 1 Select an SRS of n PSUs from the population of N PSUs

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
- ② Selection an SRS of m SSUs from the selected PSU of size M

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
- ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
- ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
- ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

Two-stage cluster sampling with equal sizes

- 1 Select an SRS of n PSUs from the population of N PSUs
- 2 Selection an SRS of m SSUs from the selected PSU of size M

- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$

Two-stage cluster sampling with equal sizes

- 1 Select an SRS of n PSUs from the population of N PSUs
- 2 Selection an SRS of m SSUs from the selected PSU of size M

- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$

- Average within-cluster variance:

$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$

Two-stage cluster sampling with equal sizes

- 1 Select an SRS of n PSUs from the population of N PSUs
- 2 Selection an SRS of m SSUs from the selected PSU of size M

- Population SSU level -

Population mean:

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$

- Population mean in PSU i :

$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$

- Population variance:

$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$

- Within PSU i variance:

$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$

- Average within-cluster variance:

$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$

- Between PSU variance:

$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
 - Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$
 - Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in S_i} \frac{M y_{ij}}{m}$$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$
 - Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in S_i} \frac{M y_{ij}}{m}$$
 - $$s_t^2 = \frac{1}{n-1} \sum_{i \in S} (\hat{t}_i - \hat{t}/n)^2$$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$
 - Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in S_i} \frac{M y_{ij}}{m}$$
 - $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (\hat{t}_i - \hat{t}/n)^2$
 - Sample element variance within PSU i : $s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m-1}$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$
 - Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in S_i} \frac{M y_{ij}}{m}$$
 - $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (\hat{t}_i - \hat{t}/n)^2$
 - Sample element variance within PSU i : $s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m-1}$
 - Avg within-cluster sampling variance: $s_w^2 = \frac{1}{n} \sum_i s_i^2$

Two-stage cluster sampling with equal sizes

- ① Select an SRS of n PSUs from the population of N PSUs
 - ② Selection an SRS of m SSUs from the selected PSU of size M
- Population SSU level -
Population mean:
$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M \frac{y_{ij}}{M_0}$$
 - Population mean in PSU i :
$$\bar{Y}_i = \sum_{j=1}^M \frac{y_{ij}}{M} = \frac{t_i}{M}$$
 - Population variance:
$$S^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(y_{ij} - \bar{Y})^2}{M_0 - 1}$$
 - Within PSU i variance:
$$S_i^2 = \sum_{j=1}^M \frac{(y_{ij} - \bar{Y}_i)^2}{M - 1}$$
 - Average within-cluster variance:
$$S_w^2 = \sum_{i=1}^N S_i^2 / N$$
 - Between PSU variance:
$$S_B^2 = \sum_{i=1}^N \frac{(\bar{Y}_i - \bar{Y})^2}{N - 1}$$
 - Sample quantities - Sample mean for PSU i : $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m}$
 - Estimated total for PSU i :
$$\hat{t}_i = \sum_{j \in S_i} \frac{M y_{ij}}{m}$$
 - $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (\hat{t}_i - \hat{t}/n)^2$
 - Sample element variance within PSU i : $s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m-1}$
 - Avg within-cluster sampling variance: $s_w^2 = \frac{1}{n} \sum_i s_i^2$
 - Between PSU variance:
$$s_B^2 = \frac{1}{n-1} \sum_{i \in S} (\bar{y}_i - \bar{y})^2$$

Estimation of sampling variance

- Sampling variance $Var(\bar{y}) = (1 - \frac{n}{N}) \frac{S_B^2}{n} + (1 - \frac{m}{M}) \frac{S_w^2}{mn}$

Estimation of sampling variance

- Sampling variance $Var(\bar{y}) = (1 - \frac{n}{N}) \frac{S_B^2}{n} + (1 - \frac{m}{M}) \frac{S_w^2}{mn}$
- The estimation $E(s_w^2) = S_w^2$ and $E(s_B^2) = S_B^2 + (1 - \frac{m}{M}) \frac{S_w^2}{m}$

Estimation of sampling variance

- Sampling variance $Var(\bar{y}) = (1 - \frac{n}{N}) \frac{S_B^2}{n} + (1 - \frac{m}{M}) \frac{S_w^2}{mn}$
- The estimation $E(s_w^2) = S_w^2$ and $E(s_B^2) = S_B^2 + (1 - \frac{m}{M}) \frac{S_w^2}{m}$
- The estimation $var(\bar{y}) = (1 - \frac{n}{N}) \frac{s_B^2}{n} + \frac{n}{N} (1 - \frac{m}{M}) \frac{s_w^2}{mn}$

Estimation of sampling variance

- Sampling variance $Var(\bar{y}) = (1 - \frac{n}{N}) \frac{S_B^2}{n} + (1 - \frac{m}{M}) \frac{S_w^2}{mn}$
- The estimation $E(s_w^2) = S_w^2$ and $E(s_B^2) = S_B^2 + (1 - \frac{m}{M}) \frac{S_w^2}{m}$
- The estimation $var(\bar{y}) = (1 - \frac{n}{N}) \frac{s_B^2}{n} + \frac{n}{N} (1 - \frac{m}{M}) \frac{s_w^2}{mn}$
- The ultimate cluster idea leads to an approximation

Ultimate clusters

- Divide each PSU into $\frac{M}{m}$ subsamples (or ultimate clusters) of m elements each
- Select SRS of a subsamples and enumerate completely
- Ultimate clusters themselves selected without replacement
- Similar to a two-stage selection but the 1st stage selection with replacement, with sampling variance:

$$var(\bar{y}) = (1 - \frac{nm}{NM}) \frac{s_B^2}{n}$$

Equal size cluster sampling: Projection

- Given an estimated design effect and a choice of M (size of each equal-size cluster), we can estimate roh (the intra-cluster correlation)
 - Example: $deff = 3.45$, $M = 50$, $roh = (3.45 - 1)/(50 - 1) = 0.05$
- Clusters are naturally occurring, and roh is a property of those clusters; we have no control over roh once we define the clusters that we wish to work with
- roh is variable-specific! Hence, design-effects are estimate-specific.
- Compute an SRS sample size and $deff$, and then compute the cluster sample size: $n_{SRS} * (1 + (\bar{M} - 1) * \hat{roh})$

Projection: roh and portability

- Given an estimate of roh from a survey (see steps and examples on previous slides), we can use that value of roh to inform new designs
- Given roh, generate a design effect based on a new choice of m (for a new /alternative sample design): $def_{new} = 1 + (m_{new}-1) * roh$
- Using the *proportion and/or element variance from the previous survey*, and the sample size for the new survey, estimate an SRS variance
- Compute the expected sampling variance of the new cluster sample given the new design effect and the new SRS variance
- roh and an estimate of the element variance are treated as portable across designs and building blocks for new designs

Example: Crime victimization survey

- Apartments in a large public housing project are sampled
- Responsible adult interviewed about victimization occurring to any member of the household
- $N = 400$ floors with exactly $M = 15$ apartments
- A SRSWR of $n = 10$ floors, and a SRS of $m = 5$ apartments on each selected floor

Example: Data

Sample floor	HH's "touched" by crime: t_i
1	4
2	4
3	5
4	5
5	3
6	1
7	0
8	1
9	2
10	1

Example: Estimation under two-stage cluster sampling

- The mean estimation $\bar{y} = \frac{\sum_i t_i}{nm} = \frac{26}{50} = 0.52$
- Sampling variance $var(\bar{y}) = (1 - \frac{nm}{NM}) \frac{s_B^2}{n} =$
 $(1 - \frac{10*5}{400*15}) \frac{1}{10} \frac{1}{5^2(10-1)} (\sum_i t_i^2 - \frac{(\sum t_i)^2}{10}) = 0.013399$
- SE: $se(\bar{y}) = \sqrt{var(\bar{y})} = 0.11575$
- 95% CI: $\bar{y} \pm t_{0.975, 10-1} se(\bar{y}) = (0.25806, 0.78194)$

Example: Projection

- ① SRS sampling variance:

$$var_{srs}(\bar{y}) = (1 - f) \frac{p(1-p)}{nm-1} = (1 - \frac{10*5}{400*15}) \frac{0.52*0.48}{50-1} = 0.005051$$

② $def f = \frac{var(\bar{y})}{var_{srs}(\bar{y})} = \frac{0.013399}{0.005051} = 2.6525$

③ $\hat{roh} = \frac{def f - 1}{m - 1} = \frac{2.6525 - 1}{5 - 1} = 0.41313$

- ④ A new design with $m_{new} = 10$ and

$$def f_{new} = 1 + (m_{new} - 1) * \hat{roh} = 1 + 9 * 0.41313 = 4.71817$$

- ⑤ New SRS sampling variance

$$var_{srs-new}(\bar{y}) = (1 - f_{new}) \frac{p(1-p)}{nm_{new}-1} = (1 - \frac{10*10}{400*15}) \frac{0.52*0.48}{100-1}$$

- ⑥ New cluster sampling variance

$$var_{cluster-new}(\bar{y}) = var_{srs-new}(\bar{y}) * def f_{new}$$

Projection: Subsample Size

- Changing subsample size (m) for fixed $m * n$ will change the number of clusters, the design effect, and the effective sample size, but not ρ_h
- As m increases, $deff$ increases (ρ_h fixed), and the effective sample size decreases
- Precision of estimates depends on n and m
- Optimum choice for n and m can be obtained by minimizing the sampling variance for fixed cost (or vice versa)
- Let C = total budget, C_0 = fixed costs subtracted from total budget (e.g., overhead, management; don't depend on sampling)

Optimum subsample Size

- Model: Sampling cost = $C - C_0 = n * c_n + n * m * c_m$ with cost per cluster (c_n) and cost per element (c_m)
- Find m to minimize the sampling variance $Var(\bar{y}) = \frac{S_B^2}{n} + \frac{S_w^2}{nm}$
- The optimum m is $m_{opt} = \sqrt{\frac{c_n}{c_m} \frac{1-roh}{roh}}$
- As roh increases, take fewer elements from each cluster

Cluster sampling: Projection

- What overall precision is needed?
- What the PSU size should be?
- How many SSUs should be sampled from each selected PSU?
- How many PSUs should be sampled?

Summary

- Cluster sampling: naturally occurring clusters (groups of elements); reduces cost but increases sampling variance
- One-stage cluster sampling
 - Variance estimation: only involves between-cluster variance
 - Design effect: portability of ρ_{hh} and element variances, projected sample size
- Two-stage cluster sampling
 - Variance estimation: involves both between- and within-cluster variance; ultimate clusters
- Portability: estimated ρ_{hh} from one survey and project future designs
- Optimum subsample size: optimum choice for n and m can be obtained by minimizing the sampling variance for fixed cost (or vice versa)

Reference

Holt, D., and T. M. F. Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society Series A* 142 (1): 33–46.