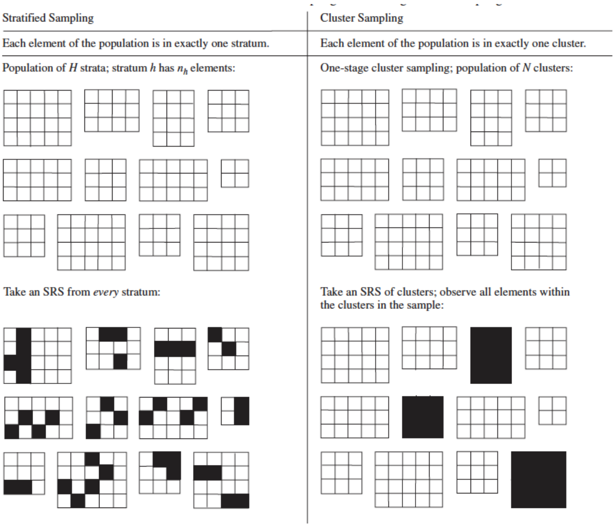
Lecture 5: Cluster sampling

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Cluster sampling



Why clustering?

- Constructing a sampling frame list of observation units may be difficult, expensive, or impossible.

  Example: We cannot list all customers of a store.

- The population may be widely distributed geographically or may occur in natural clusters such as households or schools, and it is less expensive to take a sample of clusters rather than an SRS of individuals

Whereas stratification generally increases precision when compared with simple random sampling, cluster sampling generally decreases it.

Notations

psu Level—Population Quantities

$$N = \text{number of psus in the population}$$

$$M_i = \text{number of ssus in psu } i$$

$$M_0 = \sum_{i=1}^{N} M_i = \text{total number of ssus in the population}$$

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{total in psu } i$$

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

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2 = \text{population variance of the psu totals}$$

ssu Level—Population Quantities

$$\bar{y}_U = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{y_{ij}}{M_0} = \text{ population mean}$$

$$\bar{y}_{iU} = \sum_{j=1}^{M_i}\frac{y_{ij}}{M_i} = \frac{t_i}{M_i} = \text{ population mean in psu } i$$

$$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{(y_{ij}-\bar{y}_U)^2}{M_0-1} = \text{ population variance (per ssu)}$$

$$S_i^2 = \sum_{j=1}^{M_i}\frac{(y_{ij}-\bar{y}_{iU})^2}{M_i-1} = \text{ population variance within psu } i$$

Sample Quantities

$$
\begin{aligned}
n &= \text{ number of psus in the sample} \\
m_i &= \text{ number of ssus in the sample from psu } i \\
\bar{y}_i &= \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i} = \text{ sample mean (per ssu) for psu } i \\
\hat{t}_i &= \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = \text{ estimated total for psu } i \\
\hat{t}_{\text{unb}} &= \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i = \text{ unbiased estimator of population total} \\
s_t^2 &= \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2
\end{aligned}
$$

Example: GPA

A student wants to estimate the average grade point average (GPA) in his dormitory. Instead of obtaining a listing of all students in the dorm and conducting an SRS, he notices that the dorm consists of 100 suites, each with four students; he chooses 5 of those suites at random, and asks every person in the 5 suites what her or his GPA is.

Person Number	Suite (psu) 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

$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + 8.96 + 12.96 + 11.08) = 1130.4.$$

The average of the suite totals is estimated by $\bar{t} = 1130.4/100 = 11.304$, and

$$s_t^2 = \frac{1}{5-1}\left[(12.16 - 11.304)^2 + \cdots + (11.08 - 11.304)^2\right] = 2.256.$$

Thus, $\hat{\bar{y}} = 1130.4/400 = 2.826$, and

$$\mathrm{SE}(\hat{\bar{y}}) = \sqrt{\left(1 - \frac{5}{100}\right)\frac{2.256}{(5)(4)^2}} = 0.164.$$

Population ANOVA Table—Cluster Sampling

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

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

For the GPA data, $\widehat{\text{SSB}} = (99)(0.56392) = 55.828$ and $\widehat{\text{SSW}} = (300)(0.18504) = 55.512$. Consequently, $\widehat{\text{SSTO}} = 55.828 + 55.512 = 111.340$. The estimates of the population sums of squares are given in the following table:

	df	$\widehat{\text{SS}}$ (estimated)	$\widehat{\text{MS}}$
Between suites	99	55.828	0.56392
Within suites	300	55.512	0.18504
Total	399	111.340	0.279

$$\widehat{\text{ICC}} = 1 - \frac{M}{M-1}\frac{\widehat{\text{SSW}}}{\widehat{\text{SSB}} + \widehat{\text{SSW}}} = 1 - \left(\frac{4}{3}\right)\frac{55.512}{111.34} = 0.335$$

and

$$\hat{R}_a^2 = 1 - \frac{\widehat{\text{MSW}}}{\hat{S}^2} = 1 - \frac{0.18504}{0.279} = 0.337.$$

The increase in variance for using cluster sampling is estimated to be

$$\frac{\widehat{\text{MSB}}}{\hat{S}^2} = \frac{0.56392}{0.279} = 2.02.$$

This says that we need to sample about 2.02 n elements in a cluster sample to get the same precision as an SRS of size n. There are 4 persons in each psu, so in terms of precision, one psu is worth about 4/2.02 = 1.98 SRS persons. ■

Thus, for cluster sampling,

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

If MSB/MSW is large in cluster sampling, then cluster sampling decreases precision.

How much precision do we lose by taking a cluster sample?

$$\frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = \frac{\text{MSB}}{S^2} = \frac{NM-1}{M(N-1)}[1+(M-1)\text{ICC}].$$

The ICC is only defined for clusters of equal sizes. An alternative measure of homogeneity in general populations is the adjusted R^2, called R_a^2 and defined as

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

If all psus are of the same size, then the increase in variance due to cluster sampling is

$$\frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = \frac{\text{MSB}}{S^2} = 1 + \frac{N(M-1)}{N-1}R_a^2$$

Clusters of Unequal Sizes

Unbiased Estimation. An unbiased estimator of t is calculated:

$$\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i,$$

and,

$$\text{SE}(\hat{t}_{\text{unb}}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}.$$

The difference between unequal- and equal-sized clusters is that the variation among the individual cluster totals t_i is likely to be large when the clusters have different sizes.

Clusters of Unequal Sizes

Since one-stage cluster sampling is used, an ssu is included in the sample whenever its psu is included in the sample. Thus,

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

One-stage cluster sampling produces a self-weighting sample when the psus are selected with equal probabilities. Using the weights,

$$\hat{t}_{\text{unb}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}.$$

Define

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

as the total number of ssus in the population.

Then

$$\hat{\bar{y}}_{\text{unb}} = \hat{t}_{\text{unb}}/M_0, \ \text{SE}(\hat{\bar{y}}_{\text{unb}}) = \text{SE}(\hat{t}_{\text{unb}})/M_0.$$

This estimator is inefficient for unequal sizes.

Ratio estimate

Ratio Estimation.

The population mean $\bar{y}_U$ is a ratio :

$$\bar{y}_U = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{t}{M_0},$$

where t_i and M_i are usually positively correlated. Thus, $\bar{y}_U = B$ (substituting t_i for y_i and using M_i as the auxiliary variable x_i). Define

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i \in \mathcal{S}} t_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i}.$$

Note that $\hat{\bar{y}}_r$ can also be calculated using the weights w_{ij}, as

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}.$$

Since an SRS of clusters is selected, all the weights are the same with $w_{ij} = N/n$.

Ratio estimate

$$\mathrm{SE}(\hat{\bar{y}}_r) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i \in \mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1}}$$

$$= \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i \in \mathcal{S}} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}}.$$

The variance of the ratio estimator depends on the variability of the means per element in the clusters, and can be much smaller than that of the unbiased estimator $\hat{\bar{y}}_{\text{unb}}$.

Example

Consider a population of 187 high school algebra classes in a city. An investigator takes an SRS of 12 of those classes and gives each student in the sampled classes a test about function knowledge.

Class Number	M_i	$\bar{y}_i$	t_i	$M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2$
23	20	61.5	1,230	456.7298
37	26	64.2	1,670	1,867.7428
38	24	58.4	1,402	9,929.2225
39	34	58.0	1,972	24,127.7518
41	26	58.0	1,508	14,109.3082
44	28	64.9	1,816	4,106.2808
46	19	55.2	1,048	19,825.3937
51	32	72.1	2,308	93,517.3218
58	17	58.2	989	5,574.9446
62	21	66.6	1,398	7,066.1174
106	26	62.3	1,621	33.4386
108	26	67.2	1,746	14212.7867
Total	299		18,708	194,827.0387

Example

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i} = \frac{18{,}708}{299} = 62.57.$$

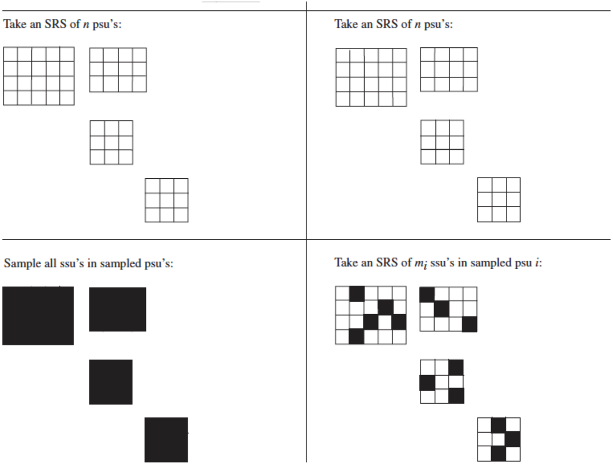
The standard error, is

$$\mathrm{SE}(\hat{\bar{y}}_r) = \sqrt{\left(1 - \frac{12}{187}\right) \frac{1}{(12)(24.92^2)} \frac{194{,}827}{11}} = 1.49.$$

The weight for each observation is $w_{ij} = 187/12 = 15.5833$; we can alternatively calculate $\hat{\bar{y}}_r$ as

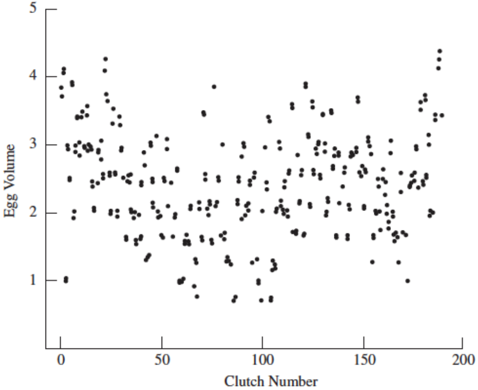
$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j=1}^{M_i} w_{ij}} = \frac{291{,}533}{4659.41667} = 62.57.$$

Two-stage cluster sampling



Example

Plot of egg volume data. Note the wide variation in the means from clutch to clutch. This indicates that eggs within the same clutch tend to be more similar than two randomly selected eggs from different clutches, and that clustering does not provide as much information per egg as would an SRS of eggs.



Example

Clutch	M_i	$\bar{y}_i$	s_i^2	$\hat{t}_i$	$(1-\frac{2}{M_i})M_i^2\frac{s_i^2}{m_i}$	$(\hat{t}_i - M_i\hat{\bar{y}}_r)^2$
1	13	3.86	0.0094	50.23594	0.671901	318.9232
2	13	4.19	0.0009	54.52438	0.065615	490.4832
3	6	0.92	0.0005	5.49750	0.005777	89.22633
4	11	3.00	0.0008	32.98168	0.039354	31.19576
⋮	⋮	⋮	⋮	⋮	⋮	⋮
182	13	4.22	0.00003	54.85854	0.002625	505.3962
183	13	4.41	0.0088	57.39262	0.630563	625.7549
184	12	3.48	0.000006	41.81168	0.000400	142.1994
sum	1757			4375.94652	42.174452	11,439.5794
$\hat{\bar{y}}_r =$		2.490579				

Example

We use the ratio estimator to estimate the mean egg volume.

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \hat{t}_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{4375.947}{1757} = 2.49.$$

From the spreadsheet (Table 5.2),

$$s_r^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - M_i \hat{\bar{y}}_r)^2 = \frac{11{,}439.58}{183} = 62.51$$

and $\bar{M}_{\mathcal{S}} = 1757/184 = 9.549$. Then,

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{9.549^2} \left[\left(1 - \frac{184}{N} \right) \frac{62.511}{184} + \frac{1}{N} \frac{42.17}{184} \right].$$

We then have

$$\mathrm{SE}(\hat{\bar{y}}_r) = \frac{1}{9.549} \sqrt{\frac{62.511}{184}} = 0.061.$$

The estimated coefficient of variation for $\hat{\bar{y}}_r$ is

$$\frac{\mathrm{SE}\,(\hat{\bar{y}}_r)}{\hat{\bar{y}}_r} = \frac{0.061}{2.49} = 0.0245.$$

Systematic sampling

Systematic sampling is a sampling plan in which the population units are collected systematically throughout the population. More specifically, a single primary sampling unit consists of secondary sampling units that are relatively spaced with each other.

Suppose we want to take a sample of size 3 from a population that has 12 elements:

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12.$$

To take a systematic sample, choose a number randomly between 1 and 4. Draw that element and every fourth element thereafter. Thus, the population contains four psus (they are clusters even though the elements are not contiguous):

$$\{1,5,9\} \quad \{2,6,10\} \quad \{3,7,11\} \quad \{4,8,12\}.$$

Now we take an SRS of one psu.