Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

Chapter 5 Cluster Sampling with Equal Probability

Example: Sampling students in high school.

- Take a random sample of n classes (The classes are the primary sampling units (psus) or clusters)
- Measure all students in the selected classes (The students within the classes are the secondary sampling units (ssus))
- Often the ssus are the elements of the population.
- In design of experiments, we would call this a nested design

Definition:

- Primary Sampling Units (PSU) or Cluster: a grouping of the members of the population, usually naturally occurring units

  Example: classes, blocks, nest of bees

- Secondary Sampling Units (SSU): units in the PSU. Often the ssu's are the elements in the population

Cluster Sampling:

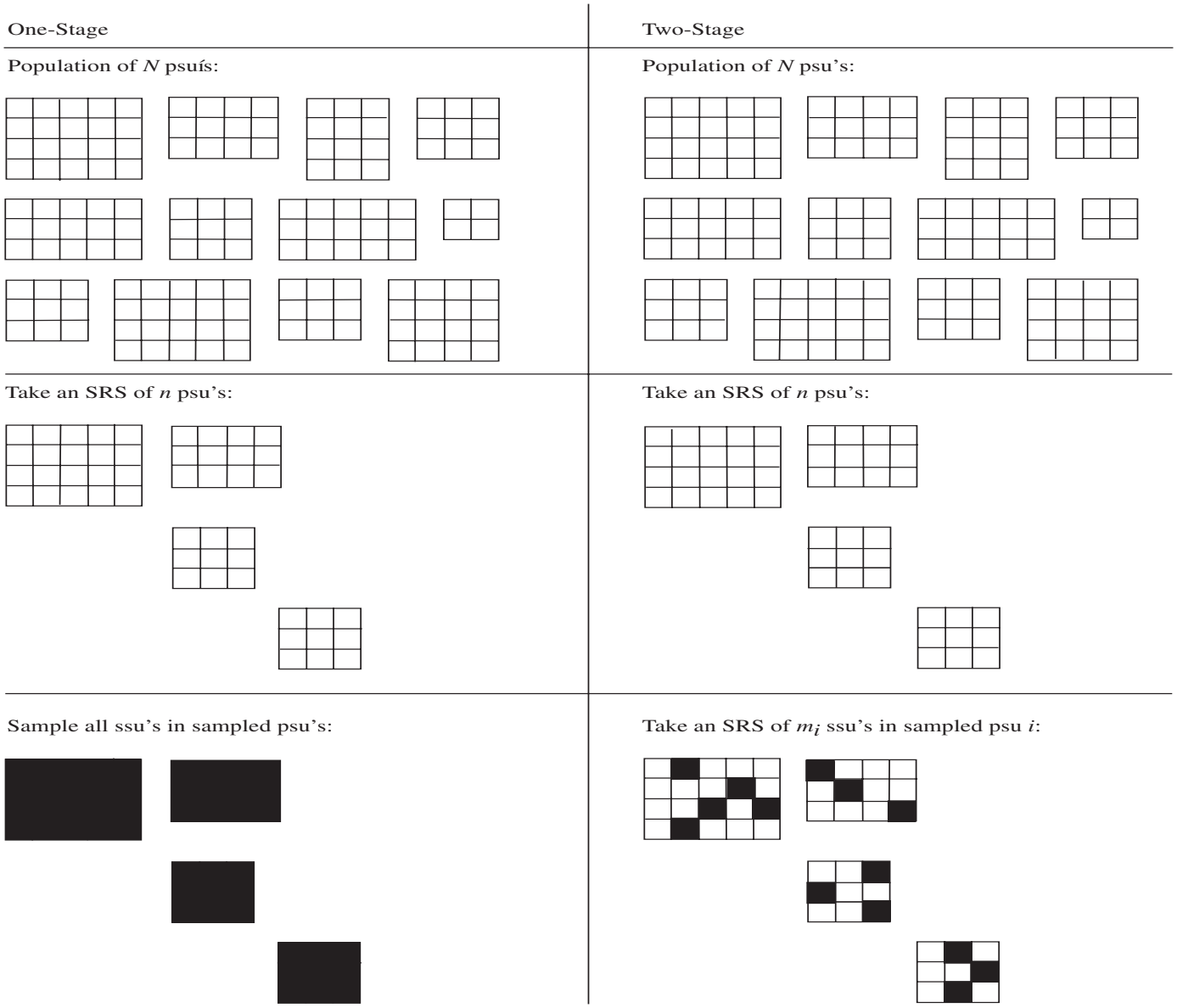
——One stage cluster sampling:

- Stage 1: Randomly select n clusters
- Stage 2: Survey all units in the selected clusters

——Two stage cluster sampling:

- Stage 1: Randomly select n clusters
- Stage 2: Survey partial of the units in the selected clusters

FIGURE 5.2
The difference between one-stage and two-stage cluster sampling.



and an unbiased estimator of the population total is

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} \hat{\tau}_i = \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \tag{5.18}$$

Comments:

- Students in the selected classes are not as likely to mirror the diversity of the high school as well as students chosen at random

  —-But it is much cheaper and easier to interview all students in the same class than students selected at random from the high school

  —- Cluster sampling may result in more information per dollar spent

- Cluster sampling complicates design and analysis and it usually decreases precision

Why use cluster sampling?

- Constructing a sampling frame list of observation units may be difficult, expensive, or impossible
  ——can't list all honeybees in a region or customers in a store
  ——possible to list all individuals in a city, but it is time-consuming and expensive, since we only have a list of housing units

- The population may be widely distributed geographically or may occur in natural clusters such as households or schools

  Example 1: Want to interview residents of nursing homes in the United States

  It is much cheaper to sample nursing homes and interview every resident in the selected homes than to interview an SRS of nursing home residents

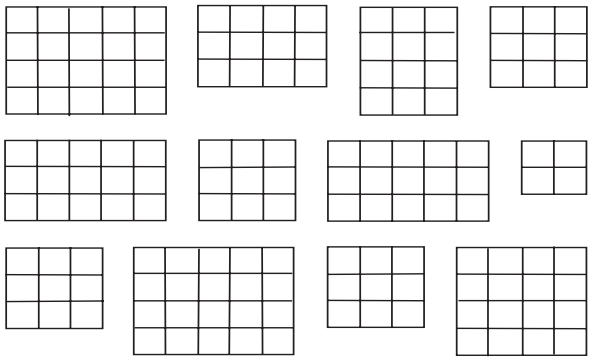
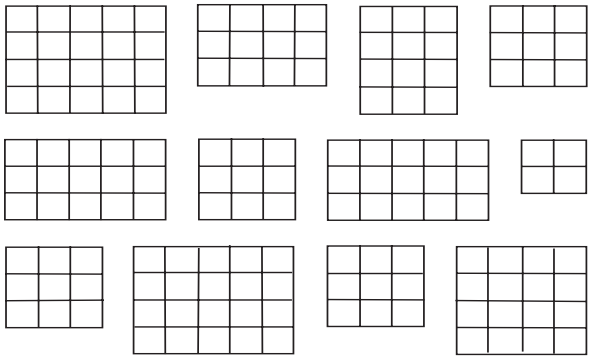
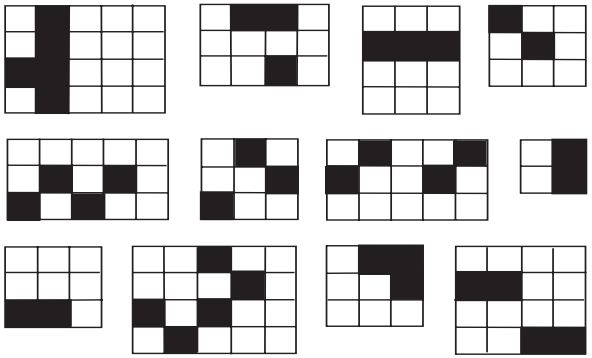
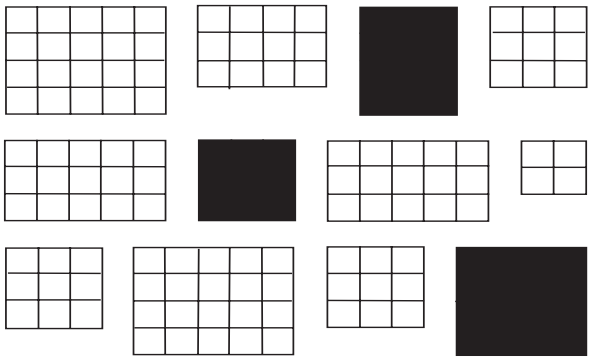
  With an SRS of residents, you might have to travel to a nursing home just to interview one resident

Example 2: In an archaeological survey, you would examine all artifacts found in a region instead of choosing points at random and examine only artifacts found at those isolated points

Comparing Cluster Sampling with Stratification

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum	Each element of the population is in exactly one cluster
population of H strata	population of N clusters
take an SRS from each stratum	take an SRS of clusters
variance of $\bar{y}_U$ depends on the variability of values within strata	variance of $\bar{y}_U$ depends primarily on the variability between cluster means
For great precision,want similar values within each stratum stratum means differ from each other	For great precision, want different values within each cluster cluster means are similar to one another

FIGURE 5.1
Similarities and differences between stratif ed sampling and one-stage cluster sampling

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata; stratum h has n_h elements: 	One-stage cluster sampling; population of N clusters: 
Take an SRS from *every* stratum: 	Take an SRS of clusters; observe all elements within the clusters in the sample: 
Variance of the estimate of $\overline{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\overline{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

Notation

y_{ij} = measurement for jth element in the ith psu

——PSU level

- N = number of psus in the population
- M_i = number of ssus in the psu i
- $M_0 = \sum_{i=1}^{N} M_i$= total number of ssus in the population
- $t_i = \sum_{j=1}^{M_i} y_{ij}$= total in psu i.
- $t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$= population total.
- $S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2$ = population variance of the psu totals (between cluster variation).

—SSU level

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

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

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

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

—Sample values

- n = number of psus in the sample
- m_i = number of elements in the sample for the ith psu
- $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i}$ = sample mean (per ssu) for psu i
- $\hat{t}_i = M_i \bar{y}_i = M_i \cdot \frac{\sum_{j \in S_i} y_{ij}}{m_i}$ = estimated total for psu i
- $\hat{t}_{\text{unb}} = N\bar{t} = N \cdot \frac{\sum_{i \in S} \hat{t}_i}{n}$ = unbiased estimator of t (population total)
- $s_t^2 = \frac{1}{n-1} \sum_{i \in S} (\hat{t}_i - \bar{t})^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2$ = estimated variance of psu totals
- $s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1}$ = sample variance within psu i

One-stage cluster sampling with equal sizes: $M_i = m_i = M$

$$\hat{t} = N\bar{t} = \frac{N}{n}\sum_{i\in S} t_i$$

$$V(\hat{t}) = N^2\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n}$$

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

S_t^2 is estimated by s_t^2 with

$$s_t^2 = \frac{1}{n-1}\sum_{i\in S}(t_i - \bar{t})^2 = \frac{1}{n-1}\sum_{i\in S}\left(t_i - \frac{\hat{t}}{N}\right)^2$$

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}$$

$$V(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{S_t^2}{nM^2}$$

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

Example: 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 a simple random sample, he notices

- the dorm consist of 100 suites, each with 4 students;
- he chooses 5 of those suites at random, and asks every person in the 5 suites what her or his GPA is.

The results are as follows:

Person	suite1	suite2	suite3	suite4	suite5
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

The psu's are the suites, $N = 100$, $n = 5$, and $M = 4$.

$$\bar{t} = (12.16 + 11.36 + 8.96 + 12.96 + 11.08)/5 = 11.304$$

$$\hat{t} = 100\bar{t} = 1130.4$$

and

$$\begin{aligned} s_t^2 &= \frac{1}{5-1}[(12.16-11.304)^2+\cdots+(11.08-11.304)^2] \\ &= 2.256 \end{aligned}$$

$$\begin{aligned} \hat{V}(\hat{t}) &= N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} \\ &= 65.4706 \end{aligned}$$

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

$$SE(\hat{\bar{y}}) = \sqrt{\left(1-\frac{5}{100}\right)\frac{2.256}{(5)(4)^2}} = .164$$

Note: Only the “total” column of the data table is used, the individual GPAs are only used for their contribution to the suite total.

ANOVA Table

Source	df	SS	MS
Between Suites	4	2.2557	.56392
Within suites	15	2.7756	.18504
Total	19	5.0313	.2648

Weight: One-stage cluster sampling with an SRS of psu's produces a self-weighting sample. The weight for each observation unit is

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

$$\begin{aligned}
\hat{t} &= \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij} \\
&= \frac{N}{n}(3.08 + 2.60 + \cdots_{+} 3.28 + 3.20) \\
&= \frac{100}{5}(56.52) \\
&= 1130.4
\end{aligned}$$

Comparing Cluster Sampling with SRS

- $S_t^2 = \sum_{i=1}^{N} \frac{(t_i - \bar{t}_U)^2}{N-1} = \sum_{i=1}^{N} \frac{M^2(\bar{y}_{iU} - \bar{y}_U)^2}{N-1} = M(MSB)$

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

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

  —MSB is relatively large since it measures the cluster-to-cluster variability: elements in different clusters often vary more than elements in the same cluster

Example: Take an SRS of classes and sampled all students for reading score within the selected class, we would likely to find that average reading scores varied from class to class

- An excellent reading teacher might raise the reading scores for the entire class
- A class of students from an area with much poverty might tend to score poorly in reading
- Unmeasured factors, such as teaching skill or poverty, can affect the overall mean for a cluster and thus cause MSB to be large
- Within a class, students' reading scores vary. If the clusters are relatively homogeneous—for example, students in the same class have similar scores- the MSW will be small.

Instead of taking a cluster sample of M elements in each of n clusters, we had taken an SRS with nM observations, the variance of the estimated total would have been

$$\begin{aligned} V(\hat{t}_{\text{srs}}) &= (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S^2}{nM} \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{MS^2}{n} \end{aligned}$$

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

If MSB $>$ S^2, cluster sampling is less efficient than simple random sampling

Intraclass Correlation Coefficient (ICC)

$$\text{ICC} = 1 - \frac{M}{M-1} \cdot \frac{\text{SSW}}{\text{SSTO}}$$

- ICC is defined to be the Pearson correlation coefficient for the $NM(M-1)$ pairs $(y_{ij}, y_{ik})$for i between 1 and N and $j \neq k$
- Defined for clusters of equal sizes
- $-\frac{1}{M-1} \leq$ ICC ≤ 1, since $0 \leq$ SSW/SSTO ≤ 1
- If the clusters are perfectly homogeneous SSW $= 0$, ICC $= 1$
- ICC tells us how similar elements in the same cluster are, or provides a measure of homogeneity within the clusters

Recall: If MSB $> S^2$, cluster sampling is less efficient than simple random sampling

- MSB $= \dfrac{NM-1}{M(N-1)} S^2[1+(M-1)\text{ICC}]$
- ICC is positive if elements within a psu tend to be similar. If the ICC is positive, cluster sampling is less efficient than simple random sampling

  —-If the clusters occur naturally in the population, ICC is usually positive

  —-Elements within the same cluster tend to be more similar than elements selected at random from the population. This may occur because the elements in a cluster share a similar environment

- ICC is negative if elements within a cluster are dispersed more than a randomly chosen group would be. This force the cluster means to be very nearly equal

  —If ICC < 0, cluster sampling is more efficient than simple random sampling of elements

  — The ICC is rarely negative in naturally occurring clusters; negative values can occur in some systematic samples or artificial clusters

Design Effect (deff)

deff(plan, statistic) = V(estimator from a sampling plan)/V(estimator from an SRS with same number of observation units)

$$\begin{aligned}\frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{srs}})} &= \frac{MSB}{S^2} \\ &= \frac{NM-1}{M(N-1)}[1+(M-1)\text{ICC}] \\ &\approx 1+(M-1)\text{ICC}\end{aligned}$$

$1+(M-1)$ICC ssus, taken in a one-stage cluster sample, give us approximately the same amount of information as one ssu from an SRS

Adjusted R^2

- An alternative quantity that can be used as a measure of homogeneity in general populations

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

- Recall: ICC $= 1 - \dfrac{M}{M-1} \cdot \dfrac{\text{SSW}}{\text{SSTO}}$
- R_a^2 is close to the ICC
- R_a^2 is the relative amount of variability in the population explained by the cluster means, adjusted for the number of degrees of freedom
- If the clusters are homogeneous, the cluster means are highly variable relative to the variation within clusters, and R_a^2 will be large

Example: Consider two artificial populations, each having three clusters with three elements per cluster

	Population A			Population B		
Cluster 1	10	20	30	9	10	11
Cluster 2	11	20	32	17	20	20
Cluster 3	9	17	31	31	32	30

- The elements are the same in the two populations, so the populations share the values $\bar{y}_U = 20$ and $S^2 = 84.5$
- In population A, most of the variability occurs within clusters; in population B, most of the variability occurs between clusters

	Population A		Population B	
	$\bar{y}_{iU}$	S_i^2	$\bar{y}_{iU}$	S_i^2
Cluster 1	20	100	10	1
Cluster 2	21	111	19	3
Cluster 3	19	124	31	1

ANOVA Table for population A:

Source	df	SS	MS	F
Between clusters	2	6	3	.03
Within clusters	6	670	111.67	
Total	8	676	84.5	

ANOVA Table for population B:

Source	df	SS	MS	F
Between clusters	2	666	333	199.8
Within clusters	6	10	1.67	
Total	8	676	84.5	

Population A:

$$R_a^2 = -.3215$$

and

$$\text{ICC} = 1 - \frac{3}{2} \cdot \frac{670}{676} = -.4867$$

- Population A has much variation among elements within the clusters but little variation among the cluster means
- Elements in the same cluster are actually less similar than randomly selected elements from the whole population
- Cluster sampling is more efficient than simple random sampling

Population B:

$$R_a^2 = .9803$$

and

$$\text{ICC} = 1 - \frac{3}{2} \cdot \frac{10}{676} = .9778$$

- Most of the variability occurs between clusters, and the clusters themselves are relatively homogeneous

- The ICC and R_a^2 are very close to 1, indicating that little new information would be gained by sampling more than one element in a cluster

- One-stage cluster sampling is much less efficient than simple random sampling

Comments:

- Most real life populations fall somewhere between the above two extremes
- The ICC is usually positive but not overly close to 1
- There is a penalty in efficiency for using cluster sampling, and that decreased efficiency should be offset by cost savings
- In general, for a given sample size, Cluster sampling will produce estimates with the largest variance. SRS will be intermediate. Stratification will give the smallest variance.

Unequal PSU Size

psu totals	t_1	t_2	$\cdots$	t_N
psu sizes	M_1	M_2	$\cdots$	M_N

Take an SRS of n psu's

Estimated population total

$$\hat{t} = N\bar{t} = \frac{N}{n}\sum_{i \in S} t_i$$

SE$(\hat{t}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n}}$, s_t^2 can be big if $t_i \propto M_i$

Example: Consider number of physicians in different areas

Number of physicians in an area usually will be proportional to the size of that area or population in that area

Population mean:

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

where t_i and M_i are usually positively correlated. Thus, $\bar{y}_U = B$ as in Section 4.1 (substituting t_i for y_i and using M_i as the auxiliary variable x_i).

- Estimate overall mean $\bar{y}_U$: $\hat{\bar{y}}_U = \hat{t}/M_0 = \hat{t}/\sum\limits_{i=1}^{N} M_i$, but $\sum\limits_{i=1}^{N} M_i$ may not be available

- Ratio estimator:

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

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum\limits_{i \in S} \sum\limits_{j \in S_i} w_{ij} y_{ij}}{\sum\limits_{i \in S} \sum\limits_{j \in S_i} w_{ij}}$$

- $$\begin{aligned}
\mathsf{SE}(\hat{\bar{y}}_r) &= \sqrt{\left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{M}^2}} \\
&= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in S}(t_i-\hat{\bar{y}}_r M_i)^2}{n-1}} \\
&= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in S}M_i^2(\bar{y}_i-\hat{\bar{y}}_r)^2}{n-1}}
\end{aligned}$$

variability depends on variability of psu means

Table 1: one stage cluster sampling

	equal size	unequal size unbiased estimator	unequal size ratio estimator
$\hat{\bar{y}}$	$\dfrac{\sum_{i\in S} t_i}{nM}$	$\dfrac{N}{n}\sum_{i\in S} t_i / \sum_{i=1}^{N} M_i$	$\sum_{i\in S} t_i / \sum_{i\in S} M_i$
$\hat{t}$	$(N/n)\sum_{i\in S} t_i$	$(N/n)\sum_{i\in S} t_i$	$M_0 \hat{\bar{y}}_r$
$V(\hat{\bar{y}})$	$(1-n/N)S_t^2/nM^2$	$N^2(1-n/N)S_t^2/nM_0^2$	$(1-n/N)S_e^2/n\bar{M}_U^2$

Notes: If all M_i's are equal, the unbiased estimator is in fact the same as the ratio estimator; If the M_i's vary, the unbiased estimator often performs poorly

Example 5.6 from textbook

One-stage cluster samples are often used in educational studies, since students are naturally clustered into classrooms or schools. 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. The (hypothetical) data are given in the file algebra.dat, with the following summary statistics.

Table 2: Example 5.6

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
$\vdots$				
108	26	67.2	1,746	14212.7867
Total	299		18,708	194,827.0387

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

$$\begin{aligned}
\mathsf{SE}(\hat{\bar{y}}_r) &= \sqrt{\left(1 - \frac{n}{N}\right) \cdot \frac{1}{n\bar{M}^2} \cdot \frac{\sum\limits_{i \in S} M_i^2 (\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}} \\
&= \sqrt{\left(1 - \frac{12}{187}\right) \cdot \frac{1}{12 \cdot 24.92^2} \cdot \frac{194,827}{11}} \\
&= 1.49
\end{aligned}$$

Two-stage cluster sampling

If the items within a cluster are very similar, no need to measure all of them. Alternative is to take an SRS of the units in each selected psu (cluster).

- First: take an SRS of n psus from the population (N psus)
- Second: for each of the sampled clusters, draw an SRS of size m_i

Need to estimate t_i,

$$\hat{t}_i = M_i \bar{y}_i, \text{ with } \bar{y}_i = \frac{1}{m_i} \sum_{j \in S_i} y_{ij}$$

The estimated total is

$$\hat{t}_{\text{unb}} = N\bar{t} = \frac{N}{n} \sum_{i \in S} \hat{t}_i = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i = \sum_{i \in S} \sum_{j \in S_i} \frac{N}{n} \frac{M_i}{m_i} y_{ij}$$

$$\begin{aligned} & p\big(\text{jth ssu in ith psu is selected}\big) \\ = \quad & p\big(\text{ith psu selected}\big) \times p\big(\text{jth ssu selected} \big| \text{ith psu selected}\big) \\ = \quad & \frac{n}{N} \cdot \frac{m_i}{M_i} \end{aligned}$$

$$\hat{t}_{\text{unb}} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}, \text{ where } w_{ij} = \frac{N M_i}{n m_i}$$

Variance for two-stage cluster sampling

The variance of $\hat{t}_{\text{unb}}$ equals the variance of $\hat{t}_{\text{unb}}$ from one-stage cluster sampling (S_t^2) plus an extra term to account for the extra variance due to estimating the $\hat{t}_i$'s rather than measuring them directly (S_i^2)

$$V(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i}$$

Estimated variance for two-stage cluster sampling

Between cluster variance

- Viewing the $\hat{t}_i$ as an SRS

$$s_t^2 = \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2 / (n-1)$$

Within cluster variance

- viewing the y_{ij} as an SRS.

- For cluster i, $s_i^2 = \dfrac{1}{m_i - 1} \displaystyle\sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2$

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$$

Summary of unbiased estimators:

- $\hat{t}_{\text{unb}} = N\bar{t} = \frac{N}{n}\sum_{i\in S}\hat{t}_i = \frac{N}{n}\sum_{i\in S}M_i\bar{y}_i = \sum_{i\in S}\sum_{j\in S_i}\frac{N}{n}\frac{M_i}{m_i}y_{ij}$

- $\hat{V}(\hat{t}_{\text{unb}}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$

- $\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{M_0}$

- $\text{SE}(\hat{\bar{y}}_{\text{unb}}) = \frac{\text{SE}(\hat{t}_{\text{unb}})}{M_0}$

Comments:

$$s_t^2 = \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2 / (n-1)$$

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$$

s_t^2 can be very large since it is affected both by variations in the unit sizes (the M_i) and by variations in the $\bar{y}_i$. If the cluster sizes are disparate, this component is large, even if the cluster means are fairly constant.

Ratio estimation

y —cluster totals t_i

x —cluster sizes M_i

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i}$$

Recall $w_{ij} = (NM_i)/(nm_i)$,

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

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in S} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right]$$

where $s_r^2 = \dfrac{\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n-1}$

and $\bar{M}$ is the average psu size

Note: for $\hat{V}(\hat{t}_{\text{unb}})$ and $\hat{V}(\hat{\bar{y}}_r)$, the second term is usually negligible compared with the first term, and most survey software packages calculate the variance using only the first term.

Example 5.8: The case of the Six-Legged Puppy

- want to estimate the average number of legs on the healthy puppies in sampled city puppy homes
- sample city has two puppy homes: Puppy Palace (PP) with 30 puppies, and Dog's Life (DL) with 10 puppies
- select one puppy home with probability $1/2$
- after the home is selected, select 2 puppies at random from the home and use $\hat{\bar{y}}_{\text{unb}}$ to estimate the average number of legs per puppy
- use ratio estimation to estimate the average number of legs per puppy

$N = 2$ puppy homes

select $n = 1$ home with prob $1/2$,

select 2 puppies at random.

Suppose PP was selected

Not surprisingly, each of the two puppies sampled has four legs, so

$$\hat{t}_{PP} = 30 \times 4 = 120$$

An unbiased estimate for the total number of puppy legs in both homes is

$$\hat{t}_{\text{unb}} = 2 \times \hat{t}_{PP} = 240$$

The mean number of legs per puppy is

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{40} = 240/40 = 6$$

Suppose now Dog's Life was selected

$$\hat{t}_{DL} = 10 \times 4 = 40$$

An unbiased estimate for the total number of puppy legs in both homes

$$\hat{t}_{\text{unb}} = 2 \times 40 = 80$$

The mean number of legs per puppy is

$$\hat{\bar{y}}_{\text{unb}} = 80/40 = 2$$

Comments:

- the estimator is mathematically unbiased: $(6+2)/2 = 4$, so averaging over all possible samples results in the right number
- $\hat{\bar{y}}_{\text{unb}}$ is unbiased, but has big variability, because M_i vary greatly (30 vs 10)

$$\begin{aligned} V(\hat{t}_{\text{unb}}) &= \left(1 - \frac{1}{2}\right) 2^2 S_t^2 + 2 \sum_{i=1}^{2} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i} \\ &= 6400 \end{aligned}$$

- when M_i's unequal, the unbiased estimators $\hat{\bar{y}}_{\text{unb}}$ are often inefficient

Ratio estimators:

- Suppose we select Puppy Place,

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{30 \times 4}{30} = 4$$

- Suppose we select Dog's Life

$$\hat{\bar{y}}_r = \frac{10 \times 4}{10} = 4$$

- Ratio estimators are the same for the two possible samples, $V(\hat{\bar{y}}_r) = 0$

Example 5.7:

- The data coots.dat come from Arnold's (1991) work on egg size and volume of American coot eggs in Minnedosa, Manitoba
- In this data set, we look at volumes of a subsample of eggs in clutches (nests of eggs) with at least two eggs available for measurement
- Randomly select 2 eggs in each clutch and measure their volume
- Want to estimate the mean egg volume

Design issues:

- what precision?
- what are the size of psus?
- how many ssus per psu?
- how many psus?

Goal of designing a survey:

- to get the most information possible for the least cost and inconvenience.

PSU size

- The psu size is often a natural unit

  Example: clutches, farms, classes, schools

- In some surveys, the investigator may have a wide choice for psu size.

  Example: want to estimate the sex and age ratios of mule deer in a region of Colorado

  psu: designed areas

  ssu: might be individual deer or groups of deer in those areas

  size of psus might be 1 km^2, 2 km^2 or 100km^2

  usually, the larger the psu size, the more variability you expect to see within a psu. Hence you expect R_a^2 and ICC to be smaller with a large psu than with a small psu. However, if psu size is too large, you may lose the cost savings of cluster sampling

Comments:

- Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.

  ——reviews optimal designs for sampling

  ——provides useful guidance for designing a survey

- There are many ways to "try out" different psu sizes before taking a survey

  —— use different combinations of R_a^2 and M, and compare the costs.

  ——-pilot study, perform an experiment and collect data on relative costs and variances with different psu sizes.

Designing a two-stage cluster survey

minimize the variance for a fixed cost

$$V(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i}$$

If $M_i = M, m_i = m$ for all psus

$$V(\hat{\bar{y}}_{\text{unb}}) = \left(1 - \frac{n}{N}\right) \frac{\text{MSB}}{nM} + \left(1 - \frac{m}{M}\right) \frac{\text{MSW}}{nm} \quad (1)$$

Choosing Subsampling Sizes

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

- a measure of homogeneity in general population
- MSW $= 0 \rightarrow R_a^2 = 1$, all elements within a cluster have the value of the cluster mean, need only subsample one element
- For other values of R_a^2, optimal allocation depends on the relative cost of sampling psus and ssus

Minimum cost

- One approach to determining sample sizes is to consider costs
- c_1 is the cost per psu; c_2 is the cost of measuring each ssu

$$\text{total cost} = C = c_1 n + c_2 n m$$

- minimize (1) to get

$$n_{\text{opt}} = \frac{C}{c_1 + c_2 m_{\text{opt}}}$$

$$m_{\text{opt}} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

Example 5.10: Recall example 5.2, A student wants to estimate the average grade point average (GPA) in his dormitory. He adopted one-stage cluster sampling plan.

- chooses 5 suites (n =5) from the 100 suites (N=100) at random
- asks every person (4 students/suite, $M_i = m_i = 4$) in the 5 suites what her or his GPA is

  Question: would subsampling have been more efficient for Example 5.2 than the one-stage cluster sample that was used?

- Set total cost $C = 300$
- Set different sets of (c_1, c_2) by $(40, 5), (10, 20)$ and $(20, 10)$
- Consider subsample size $m = (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$
  —— calculate the corresponding number of psu by $n_{\text{opt}} = C/(c_1 + c_2 m)$
  —— calculate $\hat{V}(\hat{\bar{y}}_{\text{unb}})$

- plot $\hat{V}(\hat{\bar{y}}_{\text{unb}})$ vs m

FIGURE 5.5
Estimated variance that would be obtained for the GPA example, for different values of c_1 and c_2 and different values of m. The sample estimate of 0.337 was used for R_a^2. The total cost used for this graph was $C = 300$. If it takes 40 minutes per suite and 5 minutes per person, then one-stage cluster sampling should be used; if it takes 10 minutes per suite and 20 minutes per person, then only one person should be sampled per suite; if it takes 20 minutes per suite and 10 minutes per person, the minimum is reached at $m \approx 2$, although the f atness of the curve indicates that any subsampling size would be acceptable.

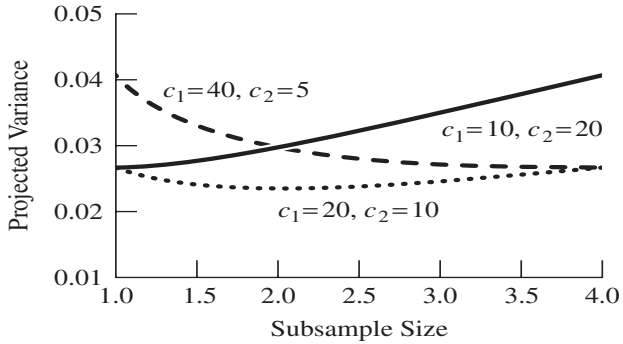
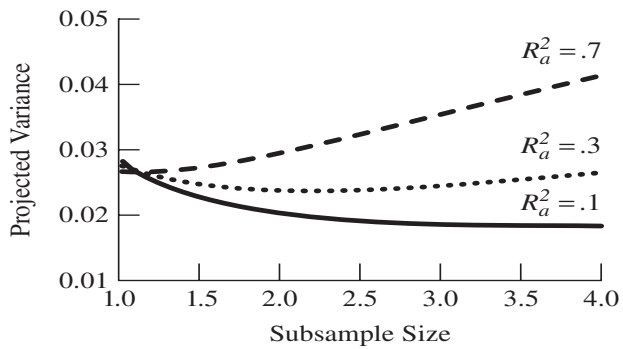


FIGURE 5.6
Estimated variance that would be obtained for the GPA example, for different values of R_a^2 and different values of m. The costs used in constructing this graph were $C = 300$, $c_1 = 20$, and $c_2 = 10$. The higher the value of R_a^2, the smaller the subsample size m should be.



information from the sample that can be used for planning future studies. Recall that $\hat{S}^2 = 0.279$, and we estimated R_a^2 by 0.337. Figures 5.5 and 5.6 show the estimated variance that would be achieved for different subsample sizes for different values of c_1 and c_2, and for different values of R_a^2. ■

Unequal PSU size (unequal M_is)

- Substitute $\bar{M}$ for M in the above work, and decide the average subsample size $\bar{m}$ to take

  —- either take $\bar{m}$ observations in every cluster

  —- or allocate observations so that $m_i/M_i =$ constant.

- As long as the M_is do not vary too much, this should produce a reasonable design

- If the M_is are widely variable, and the t_is are correlated with the M_is, a cluster sample with equal probabilities is not necessarily very efficient; an alternative design should be considered

Choosing number of psus

Assume: clusters are of equal size

$$
\begin{aligned}
V(\hat{\bar{y}}_{\text{unb}}) &= \left(1-\frac{n}{N}\right)\frac{\text{MSB}}{nM}+\left(1-\frac{m}{M}\right)\frac{\text{MSW}}{nm} \\
&\leq \frac{1}{n}\left[\frac{\text{MSB}}{M}+\left(1-\frac{m}{M}\right)\frac{\text{MSW}}{m}\right] \\
&= \frac{v}{n}
\end{aligned}
$$

An approximate $100(1-\alpha)$% CI is $\hat{\bar{y}}_{\text{unb}} \pm z_{\alpha/2}\sqrt{\frac{1}{n}v}$

If desired precision is e, then $e = z_{\alpha/2}\sqrt{\frac{1}{n}v}$

$n = z_{\alpha/2}^2 v/e^2$, v could be from a prior survey in literature

Systematic sampling:

- a special case of cluster sampling
- have a list of m units, take every kth one randomly

  Example: Want to take a systematic sample of size 3 from a population that has 12 elements: $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$
- Choose a number randomly between 1 and 4
- Draw that element and every fourth element thereafter
- The population contains $N = 4$ psus

  $S_1 = \{1, 5, 9\}, \quad S_2 = \{2, 6, 10\}$

  $S_3 = \{3, 7, 11\}, \quad S_4 = \{4, 8, 12\}$
- Take an SRS of one psu

Consider a population with NM elements

- There are N possible choices for the systematic sample, each of size M
- We observe the mean of the one cluster that comprises our systematic sample

$$\bar{y}_i = \bar{y}_{iU} = \hat{\bar{y}}_{\text{sys}}$$

Properties of $\hat{\bar{y}}_{\text{sys}}$

- $E[\hat{\bar{y}}_{\text{sys}}] = \bar{y}_U$
- For a simple systematic sample, select $n = 1$ of the N clusters

Review: One-stage cluster sampling with equal sizes: $M_i = m_i = M$

$$\hat{t} = \frac{N}{n}\sum_{i \in S} t_i, \quad V(\hat{t}) = N^2(1 - \frac{n}{N})\frac{S_t^2}{n}$$

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}, \quad V(\hat{\bar{y}}) = (1 - \frac{n}{N})\frac{S_t^2}{nM^2}$$

$$\begin{aligned} V(\hat{\bar{y}}_{\text{sys}}) &= \left(1 - \frac{1}{N}\right)\frac{S_t^2}{M^2} \\ &= \left(1 - \frac{1}{N}\right)\frac{\text{MSB}}{M} \\ &\approx \frac{S^2}{M}[1 + (M-1)\text{ICC}] \end{aligned}$$

$$V(\hat{\bar{y}}_{\text{sys}}) \approx \frac{S^2}{M}[1 + (M-1)\text{ICC}]$$

$$\text{ICC} = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}$$

- A measure of homogeneity within clusters
- ICC > 0 or large, there is little variation within the systematic samples relative to that in the population, then the elements in the sample all give similar information, systematic sampling would be expected to have higher variance than an SRS
- ICC < 0, if elements within the systematic sample (psu) are more diverse than SRS would be, systematic sampling would be more efficient than an SRS

Notes:

- Since $n = 1$ in systematic sampling, we can't obtain an unbiased estimate of $V(\hat{\bar{y}})$
- If sampling frame is in random orders, systematic sampling is a good choice
- Danger of systematic sampling, for example, the sampling frame is in the list of M,F,M,F,M,F....
- Often used when a researcher wants a representative sample of the population but does not have the resources to construct a sampling frame in advance