Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 5: Cluster Sampling with Equal Probabilities

Motivation: Sampling Students in High School

Motivation:

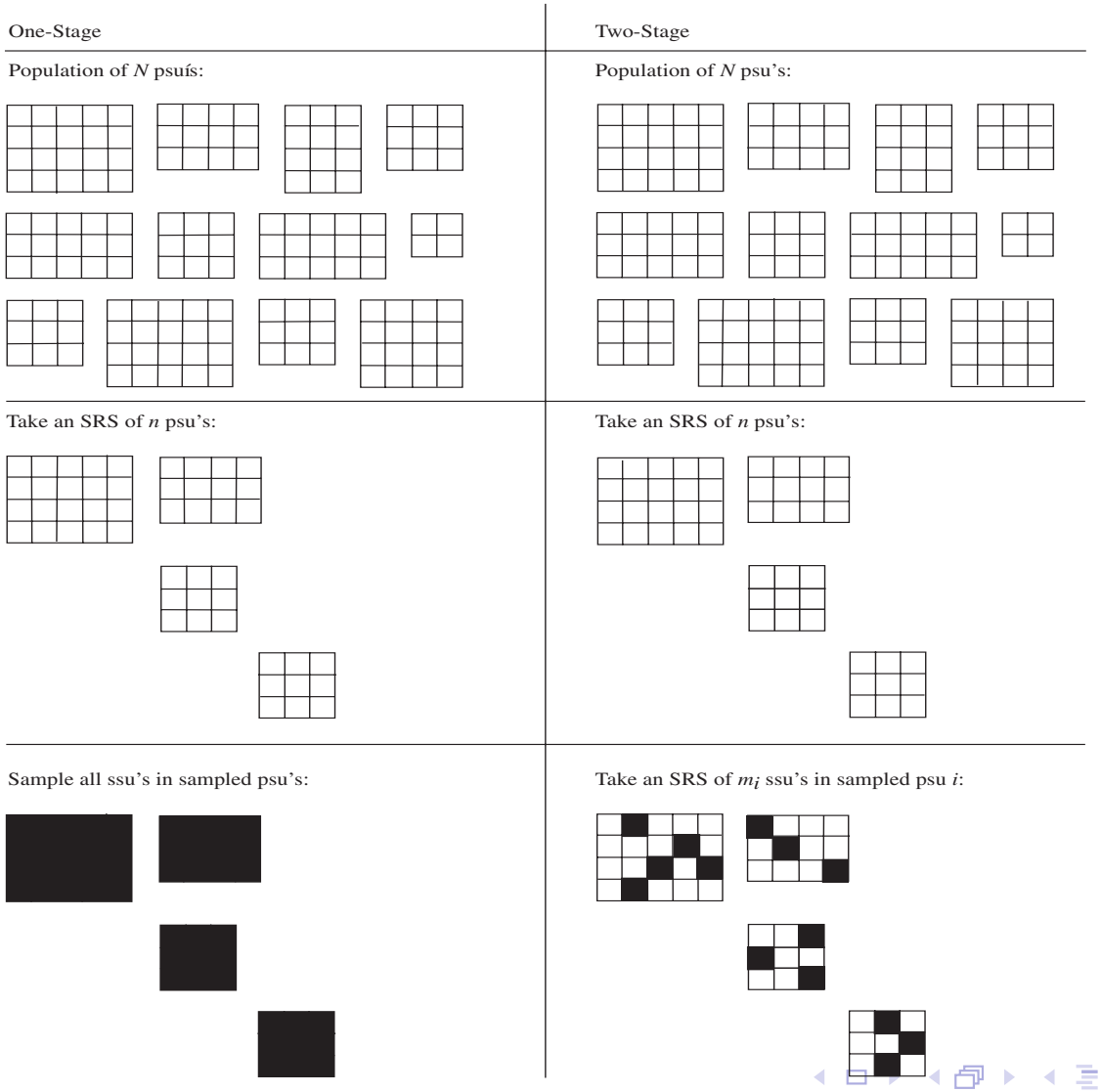
- **Simple Random Sampling (SRS):**
  - Randomly sample individual students directly from the entire school or an area.
  - This requires a complete list of all students, which can be difficult and labor-intensive to obtain.
  - Additionally, it involves locating and measuring students from various classes or schools, increasing logistical complexity and cost.
- **Cluster Sampling as an Alternative:**
  - Take a random sample of n classes (the classes are called the primary sampling units (psus) or clusters).
  - Measure all students in the selected classes (the students within the classes are called the secondary sampling units (ssus)).
  - Often, the ssus are the elements of the population.
  - In the design of experiments, this approach is called a nested design.

Cluster Sampling:

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
- 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

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 in a general case, we only have a list of housing units or a list of the phone numbers

Reasons for Using Cluster Sampling:

- Constructing a sampling frame of individual observation units can be:
  - Difficult, expensive, or even impossible.
  - For example:
    - It is impractical to list all honeybees in a region.
    - Listing all customers in a store is often unfeasible.
  - While it may be technically possible to list all individuals in a city, doing so is often:
    - Time-consuming and expensive.
    - Typically, only lists of housing units or phone numbers are available.

- 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

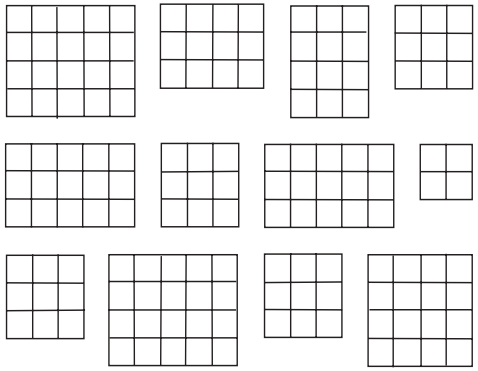
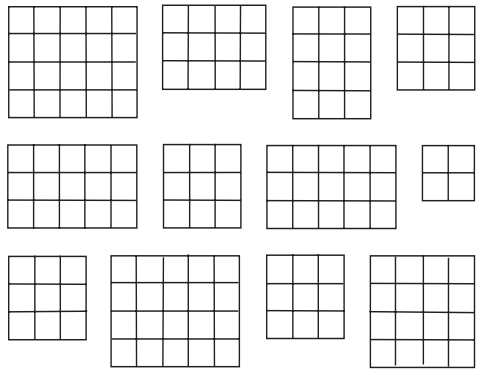
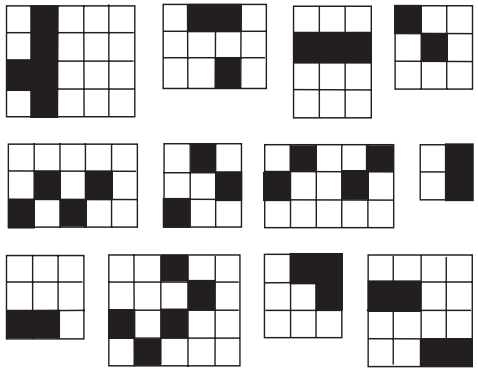
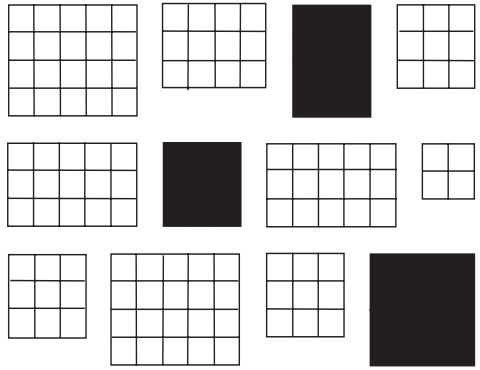
  With an SRS of residents, you might have to travel to a nursing home just to interview one resident

  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 the estimator of $\bar{y}_U$ depends on the variability of values within strata	Variance of the estimator 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 stratified 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 $\bar{y}_U$ 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

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)
- $s_t^2 = \frac{1}{n-1} \sum (\hat{t}_i - \bar{t})^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2$: 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: $M_i = m_i = M$

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

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

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

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

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

S_t^2 is estimated by s_t^2 with

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

Table 1: ANOVA Table

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 5.2: 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:

```
> print.data.frame(gpadata)
   suite  gpa wt
1      1 3.08 20
2      1 2.60 20
3      1 3.44 20
4      1 3.04 20
5      2 2.36 20
6      2 3.04 20
7      2 3.28 20
8      2 2.68 20
9      3 2.00 20
10     3 2.56 20
11     3 2.52 20
12     3 1.88 20
13     4 3.00 20
14     4 2.88 20
15     4 3.44 20
.......
20     5 3.20 20
```

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

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

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

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

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

A 95% CI for the mean is given by

$$2.826 \pm 2.776 * (0.164) = [2.37, 3.28]$$

where 2.776 is the percentile from a t distribution with $(n-1) = 4$ df. 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

```
> dgpa<-svydesign(id=~suite,weights=~wt,fpc=~rep(100,20),
data=gpadata)
> dgpa
1 - level Cluster Sampling design
With (5) clusters.
svydesign(id = ~suite, weights = ~wt, fpc = ~rep(100, 20),
data = gpadata)
> gpamean<-svymean(~gpa,dgpa)
> gpamean
     mean     SE
gpa 2.826 0.1637
> degf(dgpa)
[1] 4
> confint(gpamean,level=.95,df=4)
       2.5 %   97.5 %
gpa 2.371593 3.280407
```

```
> gpatotal<-svytotal(~gpa,dgpa)
> gpatotal
     total     SE
gpa 1130.4 65.466
> confint(gpatotal,level=.95,df=4)
       2.5 %   97.5 %
gpa 948.6374 1312.163
```

```
> suitesum<-tapply(gpadata$gpa,gpadata$suite,sum)
 #sum gpa for each suite
> suitesum
    1     2     3     4     5
12.16 11.36  8.96 12.96 11.08
> # variability comes from among the suites
> st2<-var(suitesum)
> st2
[1] 2.25568
> # SE of t-hat, formula (5.3) of SDA
> vthat <-100^2*(1-5/100)*st2/5
> sqrt(vthat)
[1] 65.46596
> # SE of ybar, formula (5.6) of SDA
> sqrt(vthat)/(4*100)
[1] 0.1636649
```

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 One-stage Cluster Sampling with SRS with nM elements

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

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

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

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

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

Intraclass Correlation Coefficient (ICC)

For clusters of equal sizes

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

- 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 $j \neq k$
- $-\frac{1}{M-1} \leq \text{ICC} \leq 1$, since $0 \leq \text{SSW}/\text{SSTO} \leq 1$
- If the clusters are perfectly homogeneous $\text{SSW} = 0$, $\text{ICC} = 1$
- 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

- $\text{MSB} = \dfrac{NM-1}{M(N-1)} S^2[1+(M-1)\text{ICC}]$
- 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 occurs 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)

$$= \frac{V(\text{estimator from a sampling plan})}{V(\text{estimator from an SRS with same number of observation units})}$$

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

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

Adjusted R^2

Adjusted R^2 is an alternative measure of homogeneity in general populations:

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

- It represents the relative amount of variability in the population explained by cluster means, adjusted for the degrees of freedom.
- When clusters are highly homogeneous:
  - Cluster means exhibit high variability relative to within-cluster variation.
  - R_a^2 will be large, reflecting greater homogeneity.
- Recall:

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

- Adjusted R^2 is a numerically adjusted approximation of the ICC, and their values are typically very close.

Example: Comparison of Two Artificial Populations
Each population contains 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

- Both populations share the same mean, $\bar{y}_U = 20$, and variance, $S^2 = 84.5$.
- In Population A, most of the variability is **within clusters**.
- In Population B, most of the variability is **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 a Simple Random Sample (SRS) of n psus.
- Estimated Population Total:

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

- Standard Error (SE):

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

- Note: s_t^2 can be large if $t_i \propto M_i$, i.e., when psu totals are proportional to psu sizes.

Example: - The number of physicians in different areas is often proportional to the area's size or population.
- Larger areas or populations tend to have more physicians, resulting in greater variability in t_i.

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 typically positively correlated. Thus, we can consider $\bar{y}_U = B$ as a ratio estimation by using M_i as the auxiliary variable.

- **Unbiased estimator** of overall mean $\bar{y}_U$:

$$\hat{\bar{y}}_U = \hat{t}_{unb}/M_0 = \hat{t}_{unb}/\sum_{i=1}^{N} M_i,$$

but $\sum_{i=1}^{N} M_i$ may not be available

- **Ratio estimator:**

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

where the denominator $\sum_{i\in S} M_i$ is a random quantity that depends on which particular psus are included in the sample.

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

- Let $e_i = t_i - M_i\hat{\bar{y}}_r = M_i(\bar{y}_i - \hat{\bar{y}}_r)$

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

$\mathrm{SE}(\hat{\bar{y}}_r)$ depends on variability between psu means

Table 2: one stage cluster sampling

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

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: One-Stage Cluster Sampling

- One-stage cluster sampling is commonly used in educational studies because students are naturally grouped into clusters such as classrooms or schools.
- Consider a population of 187 high school algebra classes ($N = 187$) in a city.
- An investigator randomly selects a simple random sample (SRS) of 12 classes ($n = 12$) and administers a test on function knowledge to all students in the selected classes.
- The (hypothetical) data for this study are provided in the file `algebra.dat`, along with the following summary statistics.

```
> nrow(algebra)
[1] 299
> head(algebra)
  class    Mi score
1    23    20    57
2    23    20    90
3    23    20    56
4    23    20    57
5    23    20    46
6    23    20    55
```

Table 3: 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
⋮				
108	26	67.2	1,746	14212.7867
Total	299		18,708	194,827.0387

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

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

A 95% CI is given by

$$62.57 \pm 2.20 * (1.49) = [59.29, 65.85],$$

where 2.20 is the percentile from a t distribution with $n-1 = 12-1 = 11$ df.

```
> algebra$sampwt<-rep(187/12,299)
> dalg<-svydesign(id=~class,weights=~sampwt,
fpc=~rep(187,299), data=algebra)
> dalg
1 - level Cluster Sampling design
With (12) clusters.

> svymean(~score,dalg)
        mean     SE
score 62.569 1.4916
> degf(dalg)
[1] 11
> confint(svymean(~score,dalg),level=.95,df=11)
         2.5 %  97.5 %
score 59.28562 65.8515
```

Two-Stage Cluster Sampling

If items within a cluster are highly similar, measuring all of them may be unnecessary. Instead, a more efficient approach is to take a simple random sample (SRS) of units within each selected primary sampling unit (psu).

- Stage 1: Select an SRS of n psus from the population of N psus.
- Stage 2: From each selected psu, draw an SRS of m_i units.

Unbiased Estimator

To estimate t_i, we use:

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

The estimated total is:

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

The probability of selecting the jth ssu within the ith psu is:

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

Thus, the unbiased estimator can also be written as:

$$\hat{t}_{\text{unb}} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}, \quad \text{where } w_{ij} = \frac{NM_i}{nm_i}.$$

Variance for Two-Stage Cluster Sampling

The variance of $\hat{t}_{\text{unb}}$ consists of two components:

- The variance from one-stage cluster sampling (S_t^2).
- An additional term to account for the extra variance (S_i^2) due to estimating the $\hat{t}_i$'s rather than measuring them directly.

The variance is expressed as:

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

Estimated Variance for Two-Stage Cluster Sampling

Between-Cluster Variance:

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

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

Within-Cluster Variance:

- Viewing the y_{ij} as an SRS.
- For cluster i,

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2$$

The estimated variance of $\hat{t}_{\text{unb}}$ is given by:

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

Summary of unbiased estimators for Two-Stage Cluster Sampling:

- $\hat{t}_{\text{unb}} = N\bar{t} = \frac{N}{n}\sum_{i\in S}\hat{t}_i = \frac{N}{n}\sum_{i\in S}M_i\bar{y}_i = \sum_{i\in S}\sum_{j\in S_i}\frac{N}{n}\frac{M_i}{m_i}y_{ij}$
- $\hat{V}(\hat{t}_{\text{unb}}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$
- $\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{M_0}$
- $\text{SE}(\hat{\bar{y}}_{\text{unb}}) = \frac{\text{SE}(\hat{t}_{\text{unb}})}{M_0}$

Recall:

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

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

Comments:

- The term s_t^2 can be very large, as it is influenced both by variations in the unit sizes (M_i) and by variations in the cluster means ($\bar{y}_i$).
- When cluster sizes vary significantly, this component becomes dominant, even if the cluster means remain relatively stable.

Ratio Estimation

Let:

- y: Cluster totals t_i
- x: Cluster sizes M_i

The ratio estimator of the mean:

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

Recall: $w_{ij} = \frac{NM_i}{nm_i}$,

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

The variance of the ratio estimator:

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

where: $s_r^2 = \dfrac{\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n-1}$ and $\bar{M} = \hat{M}_0/N$ is the average size of the psus in the sample.

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

Example 5.9: The Case of the Six-Legged Puppy

- Goal: Estimate the average number of legs on the healthy puppies in the sampled city's puppy homes.
- The sampled city has two puppy homes:
  - Puppy Palace (PP) with 30 puppies.
  - Dog's Life (DL) with 10 puppies.
- Sampling procedure:
  - Select one puppy home with probability 1/2.
  - After a home is selected, randomly select 2 puppies from that home.
- Estimation methods:
  - 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.

- Population: $N = 2$ puppy homes.
- Sampling procedure:
  - Select $n = 1$ home with probability $1/2$.
  - Randomly select 2 puppies from the selected home.

Case 1: Puppy Palace (PP) is selected

- Each of the two sampled puppies has 4 legs.
  Estimated total number of legs in PP:

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

- Unbiased estimate of the total number of puppy legs across both homes:

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

- Mean number of legs per puppy:

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

Case 2: Dog's Life (DL) is selected

- Each of the two sampled puppies has 4 legs.
  Estimated total number of legs in DL:

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

- Unbiased estimate of the total number of puppy legs across both homes:

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

- Mean number of legs per puppy:

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

Comments:

- The estimator is mathematically unbiased: $(6+2)/2 = 4$, ensuring that the average over all possible samples yields the correct value
- $\hat{\bar{y}}_{\text{unb}}$ is unbiased, but exhibits significant variability due to the considerable variation in M_i values (e.g., 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 Palace (PP):

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

- Suppose we select Dog's Life (DL):

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

- The ratio estimates are the same for the two possible samples, implying that

$$V(\hat{\bar{y}}_r) = 0.$$

Example 5.8:

- 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

```
 nrow(coots) #368
[1] 368
> head(coots)
  clutch csize length breadth volume   tmt
1      1    13   44.3    31.1  3.80     1
2      1    13   45.9    32.7  3.93     1
3      2    13   49.2    34.4  4.22     1
4      2    13   48.7    32.7  4.17     1
5      3     6   51.0    34.2 0.932     0
6      3     6   49.4    34.4 0.901     0
> coots$ssu<-rep(1:2,184)
> coots$relwt<-coots$csize/2
> dcoots<-svydesign(id=~clutch+ssu,weights=~relwt,
data=coots)
> dcoots
2 - level Cluster Sampling design (with replacement)
With (184, 368) clusters.
svydesign(id = ~clutch + ssu, weights = ~relwt,
data = coots)
```

```
svymean(~volume,dcoots)  #ratio estimator
          mean    SE
volume 2.4908 0.061
> confint(svymean(~volume,dcoots),level=.95,df=183)
           2.5 %   97.5 %
volume 2.370423 2.611134
> dcoots2<-svydesign(id=~clutch,weights=~relwt,data=coots)
> dcoots2
1 - level Cluster Sampling design (with replacement)
With (184) clusters.
> svymean(~volume,dcoots2)
          mean    SE
volume 2.4908 0.061
```

SE is the with replacement variance for the first stage sampling.

Suppose there are N clutches, weight for egg j in clutch i is

$$w_{ij} = \frac{N}{n}\frac{M_i}{m_i} = \frac{N}{184}\frac{M_i}{2}$$

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S}\sum_{j\in S_i} w_{ij}y_{ij}}{\sum_{i\in S}\sum_{j\in S_i} w_{ij}} = \frac{4375.947}{1757} = 2.49$$

$$s_r^2 = \frac{1}{n-1}\sum_{i\in S} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2 = \frac{11438.99}{183} = 62.51$$

$$\bar{M} = \sum_{i\in S} M_i/n = 1758/184 = 9.554$$

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{9.554^2}\left[\left(1 - \frac{184}{N}\right)\frac{62.51}{184} + \frac{1}{N}\frac{46.31}{184}\right]$$

Now N, the total number of clutches in the population, is unknown but presumed to be large (and known to be larger than 184).

- take the psu-level fpc to be 1
- second term in the estimated variance is small relative to the first term, and goes to 0 when $N \to \infty$

$$\hat{V}_{WR}(\hat{\bar{y}}_r) = \frac{1}{9.554^2}\frac{62.51}{184} = 0.0037,$$

which is the first-stage with replacement variance.

$$\widehat{SE}_{WR}(\hat{\bar{y}}_r) = \sqrt{0.0037} = 0.06082763$$

This is the SE reported from R without the fpc statement.

Now suppose $N = 10000$,

$$\begin{aligned}\hat{V}(\hat{\bar{y}}_r) &= \frac{1}{9.554^2}\left[\left(1-\frac{184}{N}\right)\frac{62.51}{184}+\frac{1}{N}\frac{46.31}{184}\right] \\ &= 0.003653388 + 2.757316e-07 \\ &= 0.003653663\end{aligned}$$

$$\widehat{SE}_{WOR}(\hat{\bar{y}}_r) = \sqrt{0.003653663} = 0.06044554$$

Slightly different from:

$$\widehat{SE}_{WR}(\hat{\bar{y}}_r) = \sqrt{0.0037} = 0.06082763$$

Note on Example 5.8:

- We could not use the unbiased estimator for the mean or population totals because the values of N (the total number of clusters) and $M_0 = \sum_{i=1}^{N} M_i$ (the total size of all clusters) are unknown.
- However, since the M_i's did not vary widely, the unbiased estimator would likely have had a similar coefficient of variation to the ratio estimator.
- If all the M_i's are equal, the unbiased estimator coincides with the ratio estimator.
- If the M_i's vary significantly, the unbiased estimator often performs poorly.

In *svydesign*,

- the two stages of the cluster sampling are given as id= clutch+ssu .
- when the with-replacement variance is calculated, however, as is done here, you need only specify the psus
- the point and variance estimates are the same whether you specify just the psus or you specify all stages of sampling.
- the weight argument must be included for this design because the weights are unequal.
- confidence interval uses a t critical value with 183 df (number of psus minus 1).
- svydesign does not contain the fpc argument. This is because the total number of clutches in the population, N, is unknown. —— as a result, the svymean does not use an fpc when calculating estimates.
- In general, we recommend omitting the fpc argument for multi-stage cluster sampling even when N is known.

- without fpc, it produces a variance estimate whose expectation is slightly larger than the true variance, but if n/N is small, the difference is negligible.
- if forced to choose between a standard error that is slightly too large and one that is too small, we usually prefer the former because a too-small standard error leads to claiming that estimates are more precise than they really are.
- ssus in the same psu are usually more homogeneous than randomly selected ssus from the population. Thus, the essential feature for calculating standard errors is to capture that homogeneity by including the id= psuid argument in svydesign. The issue of "to fpc or not to fpc" is minor compared with the effect of clustering.

Two-Stage Cluster Sampling: Calculate Variance with and without Replacement

Example 5.7.

The file *schools* contains data from a two-stage sample of students.

- In the first stage of sampling, an SRS of $n = 10$ schools is selected from a population of $N = 75$ schools.
- In the second stage, an SRS of $m_i = 20$ students is selected from each sampled school, and assessments for reading and math are administered.
- These data are fictional, but the summary statistics are consistent with those typically seen in educational studies.

```
> data(schools)
> print.data.frame(head(schools))
  schoolid gender math reading mathlevel readlevel  Mi finalwt
1        9      F   42      42         2         2 163  61.125
2        9      F   29      30         1         1 163  61.125
3        9      M   31      25         1         1 163  61.125
4        9      F   22      33         1         2 163  61.125
5        9      M   35      36         1         2 163  61.125
6        9      F   30      17         1         1 163  61.125
> unique(schools$schoolid)
 [1]  9 17 18 22 35 43 46 55 62 75
```

TABLE 5.7
Calculations using formulas for math scores in Example 5.7.

School	M_i	$\bar{y}_i$	s_i^2	$\hat{t}_i$	$M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$	$M_i^2(\bar{y}_i-\hat{\bar{y}}_r)^2$
9	163	34.75	74.51	5664.25	86841	70336
17	180	40.80	111.01	7344.00	159855	1909563
18	114	37.85	124.87	4314.90	66906	290396
22	367	27.95	109.31	10257.65	696046	3604196
35	109	46.10	50.31	5024.90	24401	2000806
43	219	32.20	162.80	7051.80	354749	40855
46	318	30.60	86.57	9730.80	410178	643681
55	259	36.35	141.61	9414.65	438284	698572
62	311	35.40	97.83	11009.40	442693	501495
75	263	24.60	69.52	6469.80	222134	5024481
Sum	2303			76282.15	2902087	14784382

Figure 3: Example 5.7, direct calculation for mean math scores

Source: Table 5.7 of Sampling: Design and Analysis, 3rd edition, by Sharon L. Lohr

We use the ratio estimator to estimate the mean math score. From equation (5.30),

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{76282.15}{2303} = 33.12.$$

Alternatively, we can use equation (5.31) to calculate $\hat{\bar{y}}_r$ using the sampling weights (as is done by most software packages). The weight for student j in school i is

$$w_{ij} = \frac{N}{n} \cdot \frac{M_i}{m_i} = \frac{75}{10} \cdot \frac{M_i}{20}.$$

The estimated mean math score is

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}} = \frac{572116.1}{17272.5} = 33.12.$$

With-replacement variance

The weights do not allow us to calculate the standard error directly. We need the clustering information for that. From Table 5.7,

$$s_r^2 = \frac{1}{n-1}\sum_{i\in S} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2 = \frac{14784382}{9} = 1642709.$$

Also, the average size of the sampled clusters is

$$\bar{M} = \frac{\sum_{i\in S} M_i}{n} = \frac{2303}{10} = 230.3.$$

The with-replacement variance is given by:

$$\hat{V}(\hat{\bar{y}}_r) = \frac{s_r^2}{n\bar{M}^2} = \frac{1642709}{10 \times 230.3^2} = 3.097.$$

Finally, the standard error is:

$$SE_{WR}(\hat{\bar{y}}_r) = \sqrt{3.097} = 1.76.$$

This result matches the R output provided below.

```
# 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)
```

```
mathmean<-svymean(~math,dschools)
> mathmean
       mean     SE
math 33.123 1.7599
> degf(dschools)
[1] 9
> # 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
```

Without-replacement variance

Recall that

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

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

$$\begin{aligned}\hat{V}(\hat{\bar{y}}_r) &= \frac{1}{(230.3)^2}\left[\left(1-\frac{10}{75}\right)\frac{1642709}{10} + \frac{1}{75}\frac{2902087}{10}\right]\\ &= 2.684 + 0.073\\ &= 2.757\end{aligned}$$

$$SE(\hat{\bar{y}}_r) = \sqrt{2.757} = 1.66$$

Without-replacement variance

```
# create a variable giving each student an id number
> schools$studentid<-1:(nrow(schools))
> # calculate without-replacement variance
> # 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
```

Comparison of with and without replacement variance

Variable	DF	Mean	SE	95% CI
WR	9	33.123	1.76	[29.14, 37.10]
WOR	9	33.123	1.66	[29.37, 36.88]

- The second term, 0.073, in the without-replacement variance is smaller than the first term, 2.684.
- The standard error with replacement, 1.76, is close to the standard error without replacement, 1.66.
- The confidence interval from the without-replacement method is narrower than the confidence interval from the with-replacement method.
- In general, we recommend omitting the finite population correction (fpc) argument for multi-stage cluster sampling.

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: Estimating 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 100 km^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 the 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

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

Choosing Subsampling Sizes

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

- a measure of homogeneity in general population
- $\text{MSW} = 0 \rightarrow R_a^2 = 1$, 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.
- Let c_1 be the cost of measuring each psu, and c_2 be the cost of measuring each ssu.

$$\text{Total cost} = C = c_1 n + c_2 nm$$

- Minimize equation (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, where a student wants to estimate the average grade point average (GPA) in his dormitory. He adopted a one-stage cluster sampling plan.

- Chooses 5 suites ($n = 5$) from the 100 suites ($N = 100$) at random.
- Asks every person (4 students per suite, $M_i = m_i = 4$) in the 5 suites what their GPA is.

Question: Would subsampling have been more efficient for this case than the one-stage cluster sample used in Example 5.2?

- Set total cost $C = 300$.
- Set different sets of (c_1, c_2) by $(40, 5)$, $(10, 20)$, and $(20, 10)$.
- Consider subsample sizes $m = (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$.
  - Calculate the corresponding number of psus by $n_{\text{opt}} = \frac{C}{c_1 + c_2 m}$.
  - 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 flatness of the curve indicates that any subsampling size would be acceptable.

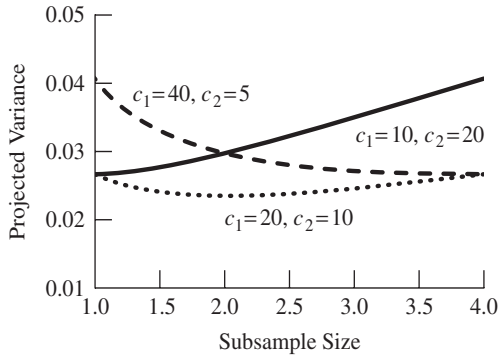
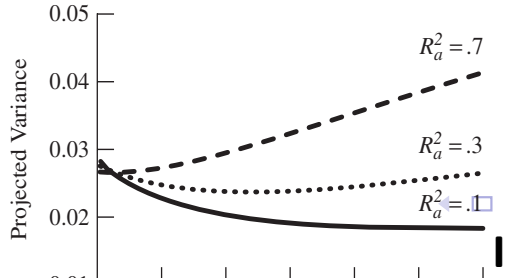


FIGURE 5.6

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.



Unequal psu Size (Unequal M_is)

- Substitute $\bar{M}$ for M in the above work, and decide the average subsample size $\bar{m}$.
  - Either take $\bar{m}$ observations in every cluster.
  - Or allocate observations so that $\frac{m_i}{M_i} = \text{constant}$.
- 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 $e = z_{\alpha/2}\sqrt{\frac{1}{n}v}$

$n = z^2_{\alpha/2}v/e^2$, 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.

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

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

- ICC is 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 cannot obtain an unbiased estimate of $V(\hat{\bar{y}})$.
- If the sampling frame is in random order, systematic sampling is a good choice.
- A potential danger of systematic sampling occurs when the sampling frame follows a regular pattern, such as Male, Female, Male, Female, Male, Female, etc.
- Systematic sampling is often used when a researcher wants a representative sample of the population but does not have the resources to construct a complete sampling frame in advance.

Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.