7 CLUSTER SAMPLING AND SYSTEMATIC SAMPLING

- In general, we want the target and study populations to be the same. When they are not the same, the researcher must be careful to ensure that conclusions based on the sample results can still be applied to the target population.

- Because of restrictions such as cost or scheduling conflicts, it is often impossible to collect a simple random sample or a stratified simple random sample. In many cases, however, it may be possible to define a sampling frame with sampling units that are not the units in the target population or the study population yet still obtain statistically valid estimates.

- *Cluster sampling*, and, specifically, *systematic sampling* are examples when a difference between the target population and the sampling frame occurs. Despite the difference, if executed properly, conclusions based on the sample results from these sampling designs can be applied to the target population.

- Situation: A population contains M_0 population units. The set of M_0 units is partitioned into N disjoint groups of population units called **primary sampling units (PSUs)**. The population units contained in the primary sampling units are called **secondary sampling units (SSUs)**.

- The primary sampling units may be of different sizes. That is, the numbers of secondary sampling units in the primary sampling units are not all the same. (See the next page.)

- Think of these disjoint groups of population units as strata. Suppose we define the sampling frame as a set of strata. Then, the sampling units in this sampling frame are not individual units in the population .

- The sampling units are **clusters** of population units. In this case, the sampling frame does not correspond with the units of the target population or the study population.

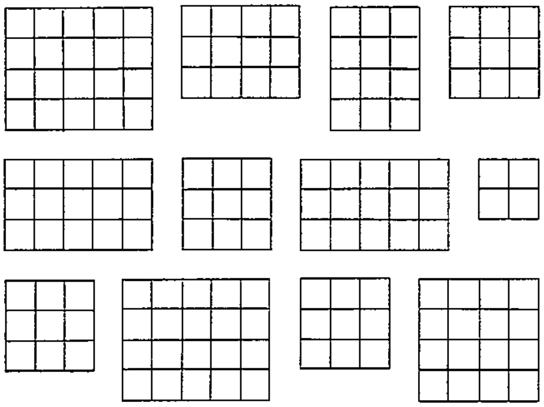
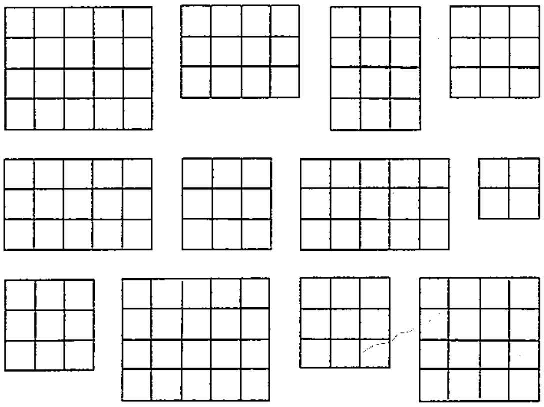
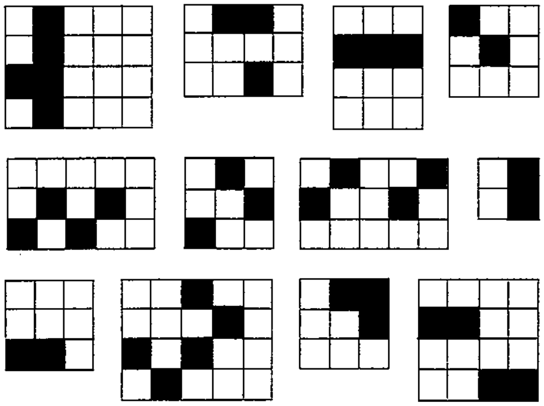
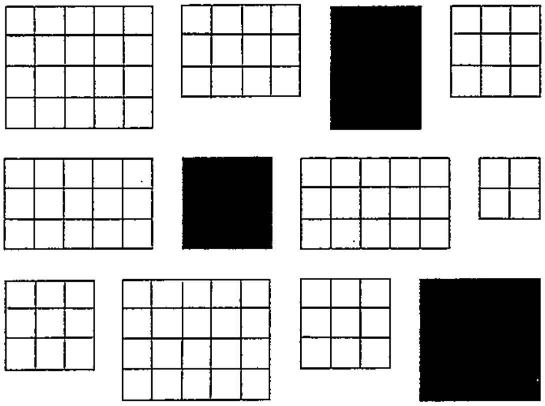
- That is, the primary sampling units (PSUs) in the sampling frame for cluster sampling are strata. Typically, the number of strata is large, while each stratum contains only a small number of secondary sampling units (SSUs).

- Note: the population has M_0 individual units but the sampling frame has only N primary sampling units corresponding the number of clusters (or strata) formed.

- The responses from the secondary sampling population units are not analyzed individually, but are combined with all other secondary sampling units that are in the same cluster. Therefore, there are N possible y values (not M_0).

- Very often, all of the secondary sampling units in each selected primary sampling unit will also be included in the sample. This is **one-stage cluster sampling** and will be studied first.

- The researcher hopes that reducing the M_0 population units to a sampling frame containing only N sampling units is offset by the practical conveniences (such as reduced cost) that this type of sampling frame can offer.

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 of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.

7.1 Notation for Cluster Sampling

The following notation will be used for each type of cluster sampling.

y_{ij} = the y-value associated with secondary sampling unit j in cluster i (SSU j in PSU i)

- **Primary sampling unit (PSU) level**

N = the number of clusters (PSUs) in the population

M_i = number of secondary sampling units (SSUs) in cluster i

$M_0 =$ $\qquad\qquad$ = the number of SSUs in the population

$t_i = \qquad\qquad\qquad =$ cluster i total $\qquad\qquad \overline{y}_i = \dfrac{t_i}{M_i} =$ cluster i mean

$t = \qquad\qquad\quad =$ population total

$$\overline{t}_i = \frac{1}{N}\sum_{i=1}^{N} t_i = \qquad\quad = \text{mean of the cluster totals (mean of PSU values)}$$

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

- **Secondary sampling unit (SSU) level**

$\overline{y}_U = \qquad\qquad =$ population mean of the SSUs

$$\overline{y}_{iU} = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} = \qquad\qquad = \text{mean of the SSUs in PSU } i$$

$$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{(y_{ij} - \overline{y}_U)^2}{M_0 - 1} = \text{population variance of SSUs.}$$

$$S_i^2 = \sum_{j=1}^{M_i}\frac{(y_{ij} - \overline{y}_{iU})^2}{M_i - 1} = \text{variance of SSUs within PSU } i.$$

- **Sample values**

$n =$ number of PSUs (clusters) in the sample

$m_i =$ number of SSUs in sampled PSU i $\quad (m_i \leq M_i)$.

$$\widehat{\overline{y}}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} = \qquad\qquad = \text{mean of the sampled SSUs in sampled PSU } i$$

$$\widehat{t}_i = \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij} = \qquad\qquad\quad = \text{estimated total of the SSUs in sampled PSU } i$$

$$\widehat{t}_{cl} = \frac{N}{n}\sum_{i=1}^{n}\widehat{t}_i = \text{unbiased estimator of population total } t$$

$$s_t^2 = \frac{\sum_{i=1}^{n}(t_i - \frac{\widehat{t}_{cl}}{N})^2}{n-1} = \text{the sample variance of estimated cluster (PSU) totals}$$

$$s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2}{m_i - 1} = \text{the sample variance within PSU } i$$

$w_{ij} =$ sampling weight for SSU j in PSU i

- Note: In one-stage cluster sampling, whe have $m_i = M_i$. That is, every SSU in PSU i is sampled. Thus, $\overline{y}_i = \widehat{\overline{y}}_i$, $t_i = \widehat{t}_i$, and $S_i^2 = s_i^2$ when $m_i = M_i$.

7.2 One-Stage Cluster Sampling

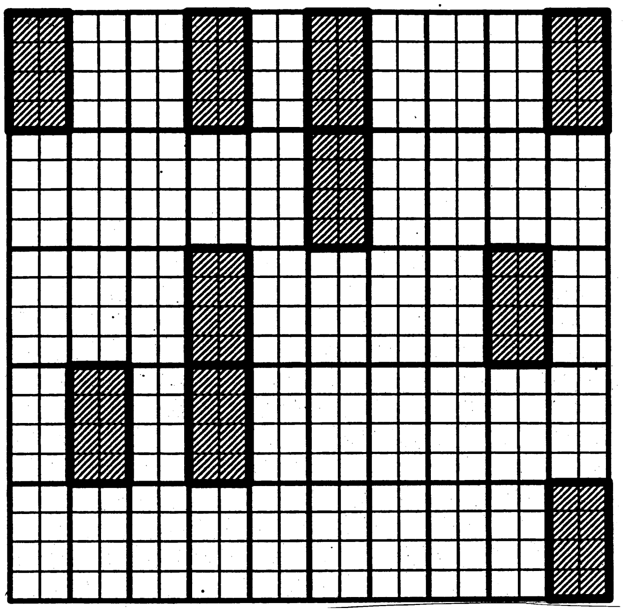
- When the strata themselves are the primary sampling units, the strata are called **clusters**. The selection of a sample of clusters to provide a sample of population units is called **cluster sampling**.

- If all of the population units in every selected cluster are in the sample, then this is known as **one-stage cluster sampling**.

- What is the difference between one-stage cluster sampling and stratified simple random sampling (stratified SRS)?
  - In stratified SRS, we take a SRS of population sampling units within each stratum to form the sample.
  - In one-stage cluster sampling, we take a subset of strata as the primary sampling units (PSUs) and then sample every SSU within each selected PSU.

- When a cluster is defined as a group of population units, the clusters are called the **primary sampling units**. Subgroups within primary sampling units are called **secondary sampling units**. For one-stage cluster sampling, the secondary sampling units are the individual population units.

A one-stage cluster sample with N = 50 clusters or primary sampling units (PSUs) each having 8 secondary sampling units (SSUs) in a population containing $M_0 = 400$ SSUs.



- If the selection of the population units within every selected cluster is restricted a second time, then this technique is known as **subsampling** or **two-stage cluster sampling**. For example, we may take a SRS of secondary sampling units within each primary sampling unit. This will be discussed later.

- If a sample of primary sampling units (Stage 1) is selected, followed by a selection of secondary sampling units (Stage 2) within the sample of primary sampling units, followed by a selection of tertiary sampling units (Stage 3) within the sample of secondary sampling units, and so on, then the sampling procedure is known as **multistage cluster sampling**.
- In cluster sampling, the size of the cluster can also be used as an auxiliary variable to select clusters with unequal sampling probabilities or used in a ratio estimator.
- Stratified sampling vs cluster sampling:
  - A researcher will use a stratified sampling design because of its potential to produce an efficient (less variable) estimator of a population characteristic. It will, in general, be more expensive to collect data for a stratified sample than for a cluster sample.
  - A researcher will use cluster sampling because of its administrative convenience. That is, cluster sampling can significantly reduce sampling costs often at the expense of a less efficient estimator of a population characteristic.

7.3 One-Stage Cluster Sampling with Equal Sized Clusters

- Suppose that each of the N clusters have the same number M of secondary sampling units ($M_1 = M_2 = \cdots = M_N = M$). Then, $M_0 = NM$.
- Suppose a SRS of n clusters (PSUs) is taken. Then the total number of SSUs selected is $m = nM$.
- There is a total of $\binom{N}{n}$ possible one-stage cluster samples and each one has the same probability of being selected. Thus, the probability of selecting any particular one-stage cluster sample $= \frac{1}{\binom{N}{n}}$.

7.3.1 Estimation of $\overline{y}_U$, t, and $\overline{t}_i$

- The unbiased estimators of $\overline{y}_U$ and t are

$$\widehat{t}_{cl} = \frac{M_0}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \frac{N}{n}\sum_{i=1}^{n} t_i = \tag{78}$$

$$\widehat{\overline{y}_{U}}_{cl} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \frac{1}{nM}\sum_{i=1}^{n} t_i = \frac{\overline{y}}{M} = \frac{N\overline{y}}{M_0} = \tag{79}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{\widehat{t}_{cl}}{N} =$ is the sample mean of the cluster (PSU) totals.

- Next, we want to study the variances of these estimators:

$$V(\widehat{t}_{cl}) = \qquad\qquad V(\widehat{\overline{y}_{U}}_{cl}) = \tag{80}$$

where $S_t^2 = \frac{\sum_{i=1}^{N}(t_i - \overline{t}_i)^2}{N-1}$ is the variance of the N cluster t_i totals. Taking a square root of the variances in (80) yields the **standard deviations** of the estimators.

- Because S_t^2 is unknown, we use the sample variance of the cluster totals:

$$s_t^2 = \frac{\sum_{i=1}^{n}(t_i - \overline{y})^2}{n-1}$$

to get unbiased estimators of the variances:

$$\widehat{V}(\widehat{t}_{cl}) \; = \qquad\qquad\qquad\qquad \widehat{V}(\widehat{\overline{y}_{Ucl}}) \; = \qquad\qquad (81)$$

- Taking the square root of the estimated variances in (81) yields the **standard errors** of the estimators.

- An unbiased estimator $(\widehat{\overline{t}_i})$ of the mean per PSU $(\overline{t}_i)$ is $\widehat{\overline{t}_i} \; = \; \overline{y} \; = \; \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} t_i \; = \; \dfrac{\widehat{t}_{cl}}{N}$.

- The variance of $\widehat{\overline{t}_i}$ is $V(\widehat{\overline{t}_i}) = \dfrac{1}{N^2}V(\widehat{t}_{cl})$ with the estimated variance being obtained by dividing the estimated variance of $\widehat{t}_{cl}$ in (81) by N^2. That is, $\widehat{V}(\widehat{\overline{y}_{U1}}) \; = \; \dfrac{N-n}{N}\dfrac{s_t^2}{n}$.

7.3.2 Confidence Intervals for $\overline{y}_U$ and t

- The confidence intervals for $\overline{y}_U$ and t are:

$$\widehat{\overline{y}_{Ucl}} \pm t^*\sqrt{\widehat{V}(\widehat{\overline{y}_{Ucl}})} \qquad\qquad \widehat{t}_{cl} \pm t^*\sqrt{\widehat{V}(\widehat{t}_{cl})} \qquad (82)$$

where t^* is the upper $\alpha/2$ critical value from the $t(n-1)$ distribution. Note that the degrees of freedom are based on n, the number of primary sampling units or sampled clusters. It is not based on the total number of secondary sampling units $m = nM$).

7.3.3 Comparison to Simple Random Sampling

- Because the variance formulas for $\widehat{\overline{y}_{Ucl}}$ and $\widehat{t}_{cl}$ in (80) are determined only from the cluster-to-cluster variability, the precision of the estimators can be improved *if clusters can be formed with small cluster-to-cluster variability.*

- We want clusters such that variability of the SSU y-values within each cluster is as large as possible but the variability of the t_i values across clusters (PSUs) is as small as possible.

- This is in contrast to stratified SRS for which we want strata such that variability within each stratum is as small as possible but the variability across strata is as large as possible.

- We will now compare $\widehat{V}(\widehat{t})$ from a SRS to $\widehat{V}(\widehat{t}_{cl})$ from a one-stage cluster sample with equal cluster sizes.

- By definition, the variance of the population of SSUs is $S^2 = \dfrac{1}{NM-1}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\overline{y}_U)^2$, and the variance of the SSUs in cluster i is $S_i^2 = \dfrac{1}{M-1}\displaystyle\sum_{j=1}^{M}(y_{ij}-\overline{y}_i)^2$. Thus,

$$\begin{aligned}
(NM-1)S^2 &= \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\overline{y}_U)^2 \quad = \quad \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\overline{y}_i+\overline{y}_i-\overline{y}_U)^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{(y_{ij}-\overline{y}_i)^2 \ + \ (\overline{y}_i-\overline{y}_U)^2 \ + \ 2(y_{ij}-\overline{y}_i)(\overline{y}_i-\overline{y}_U)\right\} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\overline{y}_i)^2 \ + \ \sum_{i=1}^{N}\sum_{j=1}^{M}(\overline{y}_i-\overline{y}_U)^2 \ + \ \sum_{i=1}^{N}\sum_{j=1}^{M}2(y_{ij}-\overline{y}_i)(\overline{y}_i-\overline{y}_U) \\
&= \sum_{i=1}^{N}(M-1)S_i^2 \ + \ \sum_{i=1}^{N}M(\overline{y}_i-\overline{y}_U)^2 \ + \ 2\sum_{i=1}^{N}(\overline{y}_i-\overline{y}_U)\left[\sum_{j=1}^{M}(y_{ij}-\overline{y}_i)\right] \\
&= (M-1)\sum_{i=1}^{N}S_i^2 \ + \ M\sum_{i=1}^{N}(\overline{y}_i-\overline{y}_U)^2 \ + \ 0 \\
&= N(M-1)\overline{S}^2 + M\sum_{i=1}^{N}(\overline{y}_i-\overline{y}_U)^2 \qquad (83)
\end{aligned}$$

where $\overline{S}^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}S_i^2$ is the average within-cluster variance. Thus, the variability of the population of SSUs can be written as a weighted sum of within-cluster and cluster-to-cluster variabilities.

- Let $N^* = NM$ and $n^* = nM$. Consider taking a SRS of n^* SSUs out of the possible N^* SSUs. That is, we to compare a SRS having the same total number of SSUs as one-stage cluster sampling. Then, $V(\widehat{t}) = \dfrac{N^*-n^*}{N^*}\dfrac{S^2}{n^*} = \dfrac{NM-nm}{NM}\dfrac{S^2}{nM}$ is the variance of the estimator of t for this SRS.

- We use (83) to compare $V(\widehat{t})$ for a SRS and $V(\widehat{t}_{cl})$ for a one-stage cluster sample . After simplification, we get:

$$V(\widehat{t}) - V(\widehat{t}_{cl}) \ = \ \frac{N^2(N-n)(M-1)}{nM(N-1)}\left(\overline{S}^2 - S^2\right) \qquad (84)$$

- If $V(\widehat{t}) - V(\widehat{t}_{cl}) > 0$ (or, if $\overline{S}^2 > S^2$), then we say that $\widehat{t}_{cl}$ **is more efficient** than $\widehat{t}$ for estimating t. This result is also true for estimation of $\overline{y}_U$. That is, if $V(\widehat{\overline{y}_U}) - V(\widehat{\overline{y}_{U cl}}) > 0$, then the one-stage cluster sample estimator $\widehat{\overline{y}_{U cl}}$ would be more efficient than SRS estimator $\widehat{\overline{y}_U}$ for estimating $\overline{y}_U$.

- Practically speaking, the one-stage cluster sample estimator will be more efficient than the SRS estimator of t or $\overline{y}_U$ if the average within-cluster variability ($\overline{S}^2$) is larger than the population variance (S^2).

Figure 7: Cluster Sampling Example for the Longleaf Pine Data

The total abundance $t = 584$. There are $M_0 = 400$ SSUs and $N = 100$ PSUs (clusters) of size $M = 4$.

1	1	1	1	1	2	1	0	0	0	4	5	0	1	0	1	2	1	0	1
3	2	1	0	1	0	0	0	1	2	2	2	0	2	2	2	0	2	0	1
7	4	1	1	1	1	0	0	0	2	2	0	4	3	2	4	2	1	2	2
0	1	2	0	0	0	0	0	4	6	5	1	5	0	0	0	2	1	2	0
1	1	0	2	3	2	0	0	2	1	3	1	4	1	1	1	2	2	1	1
2	0	0	0	4	3	3	0	1	16	5	0	1	3	8	0	0	1	3	3
0	0	1	14	3	3	1	2	0	8	0	2	0	3	9	0	4	2	1	0
0	0	5	1	8	7	6	6	6	1	0	4	0	0	1	2	2	0	1	2
0	0	2	2	3	2	2	3	1	1	1	3	0	0	2	2	0	3	4	0
0	0	0	0	1	0	3	1	1	1	2	0	2	0	2	0	2	1	1	0
1	8	7	7	8	0	5	0	1	0	1	2	0	0	2	4	2	2	2	4
0	9	1	0	0	1	1	1	0	0	0	1	2	4	0	2	1	3	3	1
0	0	0	1	0	2	4	3	1	2	2	0	0	1	1	2	2	0	2	4
0	1	0	0	1	2	0	2	3	5	2	0	0	2	1	1	2	0	1	3
1	0	0	1	1	0	0	0	2	2	2	1	1	1	0	0	2	0	0	0
0	2	0	2	2	0	1	1	0	2	0	0	1	0	0	1	1	1	5	3
0	0	0	3	2	1	0	0	0	0	0	2	1	0	1	1	1	3	1	2
1	0	0	1	0	3	0	1	0	0	2	1	2	0	0	0	1	1	1	0
0	0	0	0	0	0	0	1	1	1	0	1	0	3	0	2	0	1	1	0
2	0	0	0	0	0	0	0	1	2	0	1	3	0	0	1	0	1	2	4

The following figure contains the 100 cluster totals (t_i for sampling unit $i = 1, 2, \ldots, 100$). A SRS will be taken from these 100 t_i values.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
11	8	5	2	3	3	1	0	5	10	13	8	9	6	4	7	6	5	4	4
21	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*
3	1	6	17	18	15	10	8	9	26	8	7	5	7	19	3	8	5	6	6
41	*42*	*43*	*44*	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*	*53*	*54*	*55*	*56*	*57*	*58*	*59*	*60*
1	17	10	9	12	3	11	5	3	2	4	6	4	4	6	8	5	9	10	5
61	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*	*70*	*71*	*72*	*73*	*74*	*75*	*76*	*77*	*78*	*79*	*80*
1	3	0	4	4	4	5	6	6	11	6	1	2	4	2	4	7	1	8	10
81	*82*	*83*	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*96*	*97*	*98*	*99*	*100*
3	0	0	4	2	4	0	2	2	3	2	5	6	3	1	4	2	6	5	6

The sample contains $n = 8$ clusters (PSUs). The t_i totals for the 8 sampled PSUs are in ( )

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
11	**(8)**	5	2	3	3	1	0	5	10	13	8	9	6	4	7	6	5	4	4
21	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*
3	1	6	17	18	15	10	**(8)**	9	**(26)**	8	7	5	7	19	3	8	5	6	6
41	*42*	*43*	*44*	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*	*53*	*54*	*55*	*56*	*57*	*58*	*59*	*60*
1	17	10	**(9)**	12	3	11	5	3	2	4	6	4	4	**(6)**	8	5	9	10	5
61	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*	*70*	*71*	*72*	*73*	*74*	*75*	*76*	*77*	*78*	*79*	*80*
1	3	0	4	4	**(4)**	5	6	6	11	6	1	2	4	2	**(4)**	7	**(1)**	8	10
81	*82*	*83*	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*96*	*97*	*98*	*99*	*100*
3	0	0	4	2	4	0	2	2	3	2	5	6	3	1	4	2	6	5	6

Figure 8a: Cluster Sampling Example for a Spatially Correlated Population

The abundance counts show a strong diagonal spatial correlation. The total abundance $t = 13354$. There are $M_0 = 400$ SSUs and $N = 50$ PSUs of size $M = 8$.

18	20	15	20	20	15	19	18	24	23	20	26	29	28	28	31	31	34	28	32
13	20	16	20	15	23	19	26	21	21	24	30	23	26	25	33	31	28	32	38
16	18	20	24	25	26	22	23	26	26	22	27	25	25	34	28	37	36	38	31
17	17	16	22	21	23	22	27	27	24	28	32	29	33	27	37	37	38	35	33
15	19	23	17	21	23	21	23	24	25	31	26	32	34	32	33	31	31	36	37
21	24	20	21	28	26	30	22	31	25	29	29	27	30	29	37	35	32	38	43
23	17	24	25	24	27	31	29	31	34	27	36	29	29	34	39	37	37	40	36
18	24	21	25	27	22	32	32	31	26	28	34	34	37	35	34	38	38	37	40
22	26	28	26	24	29	33	26	27	27	34	31	39	32	36	38	37	40	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	40	37	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	40	44	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	42	41	40	49
28	24	28	27	26	31	32	29	32	33	38	34	39	38	40	37	41	43	42	43
32	25	31	32	29	29	35	38	38	32	36	35	39	42	39	40	44	42	41	45
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	49	48	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	41	43	44	51
28	35	35	34	34	33	41	33	34	35	39	44	44	48	44	50	49	48	53	54
29	33	32	36	39	33	33	34	35	42	46	47	48	47	46	45	44	52	54	55
28	37	38	37	33	33	34	37	45	40	39	42	42	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53

The following figure contains the 50 cluster totals (t_i for sampling unit $i = 1, 2, \ldots, 50$). A SRS will be taken from these 50 t_i values.

1 139	*2* 153	*3* 168	*4* 176	*5* 192	*6* 209	*7* 218	*8* 243	*9* 272	*10* 267
11 161	*12* 176	*13* 198	*14* 220	*15* 227	*16* 240	*17* 252	*18* 273	*19* 279	*20* 307
21 197	*22* 222	*23* 229	*24* 246	*25* 258	*26* 268	*27* 285	*28* 308	*29* 321	*30* 346
31 224	*32* 245	*33* 248	*34* 279	*35* 285	*36* 298	*37* 329	*38* 336	*39* 351	*40* 366
41 267	*42* 288	*43* 277	*44* 296	*45* 312	*46* 343	*47* 370	*48* 376	*49* 395	*50* 419

Figure 8b: Cluster Sampling Example for a Spatially Correlated Population

The abundance counts show a strong diagonal spatial correlation. The total abundance $t = 13354$. There are $M_0 = 400$ SSUs and $N = 100$ PSUs (clusters) of size $M = 4$.

18	20	15	20	20	15	19	18	24	23	20	26	29	28	28	31	31	34	28	32
13	20	16	20	15	23	19	26	21	21	24	30	23	26	25	33	31	28	32	38
16	18	20	24	25	26	22	23	26	26	22	27	25	25	34	28	37	36	38	31
17	17	16	22	21	23	22	27	27	24	28	32	29	33	27	37	37	38	35	33
15	19	23	17	21	23	21	23	24	25	31	26	32	34	32	33	31	31	36	37
21	24	20	21	28	26	30	22	31	25	29	29	27	30	29	37	35	32	38	43
23	17	24	25	24	27	31	29	31	34	27	36	29	29	34	39	37	37	40	36
18	24	21	25	27	22	32	32	31	26	28	34	34	37	35	34	38	38	37	40
22	26	28	26	24	29	33	26	27	27	34	31	39	32	36	38	37	40	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	40	37	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	40	44	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	42	41	40	49
28	24	28	27	26	31	32	29	32	33	38	34	39	38	40	37	41	43	42	43
32	25	31	32	29	29	35	38	38	32	36	35	39	42	39	40	44	42	41	45
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	49	48	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	41	43	44	51
28	35	35	34	34	33	41	33	34	35	39	44	44	48	44	50	49	48	53	54
29	33	32	36	39	33	33	34	35	42	46	47	48	47	46	45	44	52	54	55
28	37	38	37	33	33	34	37	45	40	39	42	42	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53

The following figure contains the 100 cluster totals (t_i for sampling unit $i = 1, 2, \ldots, 100$).

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
64	75	67	86	81	87	82	94	98	94	94	115	106	112	114	129	136	136	133	134
21	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*
77	84	88	88	100	98	114	106	117	110	115	125	122	130	130	143	141	138	151	156
41	*42*	*43*	*44*	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*	*53*	*54*	*55*	*56*	*57*	*58*	*59*	*60*
94	103	118	104	110	119	123	123	124	134	132	136	146	139	152	156	159	162	170	176
61	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*	*70*	*71*	*72*	*73*	*74*	*75*	*76*	*77*	*78*	*79*	*80*
117	107	126	119	123	125	134	145	147	138	147	151	162	167	171	165	175	176	178	188
81	*82*	*83*	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*96*	*97*	*98*	*99*	*100*
123	144	144	144	140	137	147	149	153	159	169	174	178	192	183	193	197	198	204	215

The sample contains $n = 10$ clusters (PSUs). The t_i totals for the 10 sampled PSUs are in ( )

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
64	75	67	86	**(81)**	87	82	94	98	94	94	115	106	112	114	129	**(136)**	136	133	134
21	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*
77	**(84)**	88	88	100	98	114	106	117	110	115	125	122	130	130	143	141	138	151	156
41	*42*	*43*	*44*	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*	*53*	*54*	*55*	*56*	*57*	*58*	*59*	*60*
94	103	118	104	110	119	**(123)**	123	124	134	132	136	146	**(139)**	152	156	159	162	170	176
61	*62*	*63*	*64*	*65*	*66*	*67*	*68*	*69*	*70*	*71*	*72*	*73*	*74*	*75*	*76*	*77*	*78*	*79*	*80*
117	107	126	119	123	125	134	**(145)**	147	138	147	151	162	**(167)**	171	165	**(175)**	176	178	188
81	*82*	*83*	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*96*	*97*	*98*	*99*	*100*
123	144	144	144	**(140)**	137	147	149	153	159	169	174	178	192	183	**(193)**	**(197)**	198	204	215

7.3.4 Using R and SAS for One-Stage Cluster Sampling

R code for Cluster Sample in Figure 7

- Enter the y-values of the SSUs in one vector (e.g., 'trees') and the cluster number in a second vector (e.g., 'clusterid').

```
library(survey)
source("c:/courses/st446/rcode/confintt.r")

# One-stage cluster sample from Figure 7
N =100
n =8
M =4
wgt = N/n

trees <- c(1,2,4,1,2,0,7,0,2,2,0,0,0,0,2,6,1,16,8,1,2,2,2,0,2,1,0,1,0,0,0,1)

clusterid <- c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,6,6,6,6,
7,7,7,7,8,8,8,8)

fpc <- c(rep(N,n*M))

Fig7 <- data.frame(cbind(clusterid,trees,fpc))

dsgn7 <- svydesign(ids=~clusterid,weights=c(rep(wgt,n*M)),fpc=~fpc,data=Fig7)

esttotal <- svytotal(~trees,design=dsgn7)
print(esttotal,digits=15)
confint.t(esttotal,level=.95,tdf=n-1)

estmean <- svymean(~trees,design=dsgn7)
print(estmean,digits=15)
confint.t(estmean,level=.95,tdf=n-1)
```

R output for Cluster Sample in Figure 7

```
      total     SE
trees   825 259.37

-------------------------------------------------------------------
mean( trees ) = 825.00000
SE( trees ) = 259.37425

Two-Tailed CI for trees where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  211.67737     1438.32263
-------------------------------------------------------------------

        mean     SE
trees 2.0625 0.6484

-------------------------------------------------------------------
mean( trees ) = 2.06250
SE( trees ) = 0.64844
Two-Tailed CI for trees where alpha = 0.05 with 7 df
    2.5 %        97.5 %
  0.52919     3.59581
-------------------------------------------------------------------
```

R code for Cluster Sample in Figure 8b

```
N =100
n =10
M =4
wgt = N/n

y <- c(20,15,25,21,31,31,37,37,19,24,17,24,33,25,31,34,32,32,37,38,
29,38,40,38,41,44,49,41,34,39,33,34,50,45,48,50,49,44,52,52)

clusterid <- c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,6,6,6,6,
7,7,7,7,8,8,8,8,9,9,9,9,10,10,10,10)

(The remainder of the code is the same as the previous example)
```

R output for Cluster Sample in Figure 8b

```
  total     SE
y 14130 1192.2

------------------------------------------------------------------
mean( y ) = 14130.00000
SE( y ) = 1192.22900

Two-Tailed CI for y where alpha = 0.05 with 9 df
    2.5 %        97.5 %
  11432.99062     16827.00938


    mean     SE
y 35.325 2.9806

------------------------------------------------------------------
mean( y ) = 35.32500
SE( y ) = 2.98057

Two-Tailed CI for y where alpha = 0.05 with 9 df
    2.5 %        97.5 %
  28.58248     42.06752
------------------------------------------------------------------
```

SAS code for Cluster Sample in Figure 7 (Supplemental)

- To use Proc Surveymeans to analyze data from a one-stage cluster sample with the goal of estimating $\overline{y}_U$ or t, we need to include a **Cluster** statement followed by a cluster label. In the first example, the clusters are labeled “_cluster”.
- The value following “**total =**” is the number of primary sampling units in the population.
- The appropriate weight to use in the **weight** statement to get the correct estimates for t is $M_0/(nM)$.

SAS code for Cluster Sample in Figure 7

```
DATA cluster1;
  M0 = 400;     * number of secondary sampling units (SSUs) in population;
  n  = 8;       * number of primary sampling units (PSUs) sampled;
  m  = 4;       * number of SSUs in a PSU;

  wgt = M0/(n*m);
  DO psu = 1 to n;
  DO ssu = 1 to m;
     INPUT trees @@; OUTPUT;
  END; END;
DATALINES;
1 2 4 1   2 0 7 0   2 2 0 0   0 0 2 6   1 16 8 1  2 2 2 0   2 1 0 1   0 0 0 1
;
*** TOTAL = number of PSUs in the population ***;

PROC SURVEYMEANS DATA=cluster1 TOTAL=100 MEAN CLM SUM CLSUM;
     VAR trees;
     CLUSTER psu;
     WEIGHT wgt;
TITLE 'One-Stage Cluster Sample from Figure 7';
RUN;
```

SAS output for Cluster Sample in Figure 7

```
The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                 8
Number of Observations            32
Sum of Weights                   400

                           Statistics

                               Std Error
Variable             Mean        of Mean       95% CL for Mean
-----------------------------------------------------------------
trees            2.062500       0.648436    0.52919341 3.59580659
-----------------------------------------------------------------

                           Statistics

Variable              Sum        Std Dev        95% CL for Sum
-----------------------------------------------------------------
trees          825.000000     259.374247    211.677365 1438.32263
-----------------------------------------------------------------
```

SAS code for Cluster Sample in Figure 8b

```
DATA cluster2;
  M0 = 400;     * number of secondary sampling units (SSUs) in population;
  n  = 10;      * number of primary sampling units (PSUs) sampled;
  m  = 4;       * number of SSUs in a PSU;

  wgt = M0/(n*m);
  DO psu = 1 to n;
  DO ssu = 1 to m;
     INPUT y @@; OUTPUT;
  END; END;
DATALINES;
20 15 25 21 31 31 37 37 19 24 17 24 33 25 31 34 32 32 37 38
29 38 40 38 41 44 49 41 34 39 33 34 50 45 48 50 49 44 52 52
;

*** TOTAL = number of PSUs in the population ***;

PROC SURVEYMEANS DATA=cluster2 TOTAL=100 MEAN CLM SUM CLSUM;
     VAR y;
     CLUSTER psu;
     WEIGHT wgt;
TITLE 'One-Stage Cluster Sample from Figure 8b';
RUN;
```

SAS output for Cluster Sample in Figure 8b

```
The SURVEYMEANS Procedure

            Data Summary

Number of Clusters                10
Number of Observations            40
Sum of Weights                   400

                           Statistics

                               Std Error
Variable            Mean         of Mean       95% CL for Mean
-----------------------------------------------------------------
y              35.325000        2.980573    28.5824765 42.0675235
-----------------------------------------------------------------

                           Statistics

Variable             Sum         Std Dev        95% CL for Sum
-----------------------------------------------------------------
y                  14130     1192.229005    11432.9906 16827.0094
-----------------------------------------------------------------
```