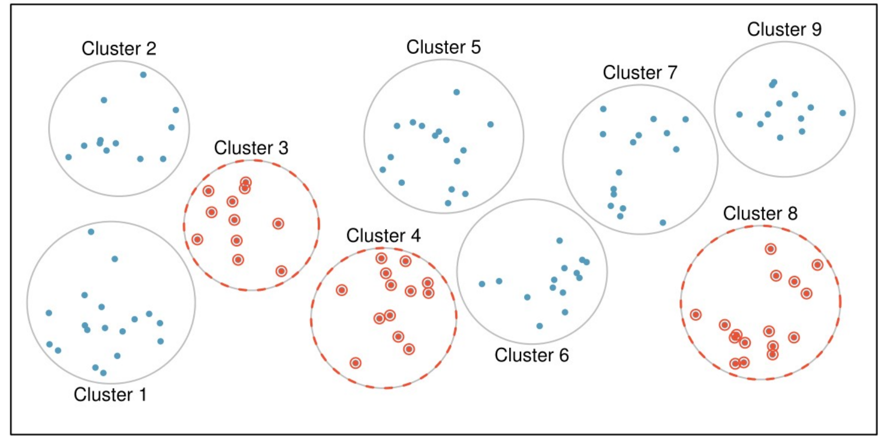
One-stage cluster sampling estimation

Week 6 (5.1, 5.2.1, 5.2.3)

Stat 260, St. Clair

Design: One-Stage Cluster Sample

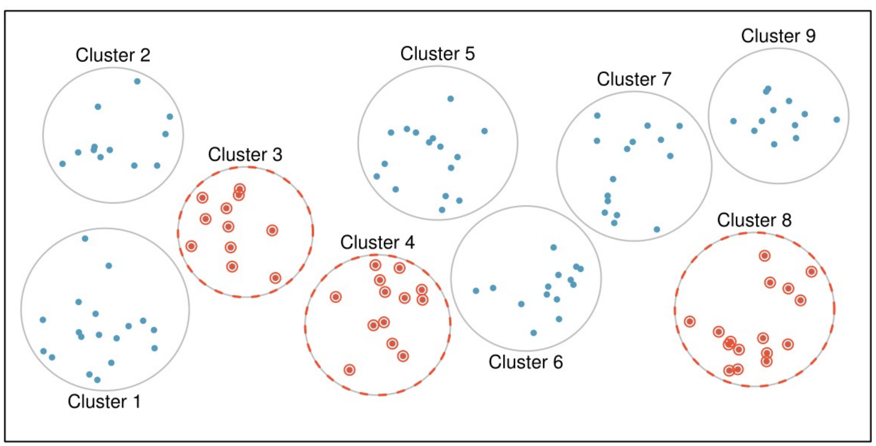
Definition: Divide all population **observation units** into N non-overlapping **clusters** of observation units.



https://spot.pcc.edu/~evega/section-4.html

Design: One-Stage Cluster Sample

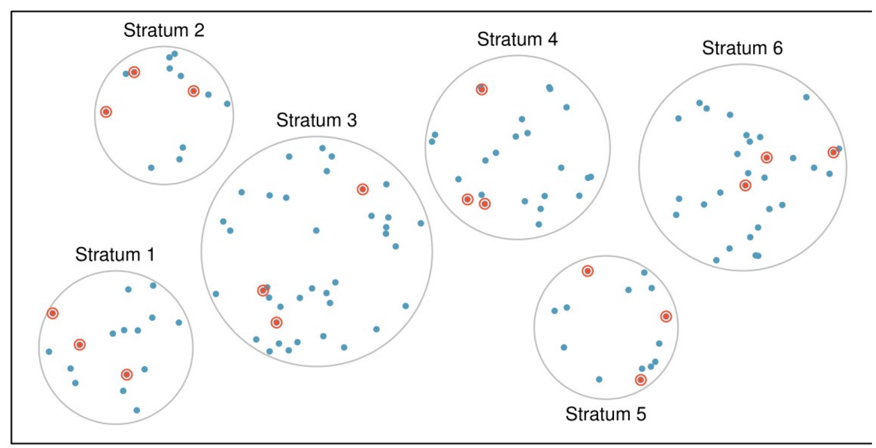
Defined: We take a SRS of n **clusters** and survey **every observation unit** in selected clusters.



https://spot.pcc.edu/~evega/section-4.html

Design: Cluster vs. Stratified sampling

Take a SRS **within** each strata



https://spot.pcc.edu/~evega/section-4.html

Design: One-Stage Cluster Sample

- **Primary Sampling Units (PSU):** clusters
- **Secondary Sampling Units (SSU):** observation units
  - y_{ij} is the measurement for unit j in cluster i
  - M_i is the number of observation units in cluster i
  - $M_0 = \sum_{i=1}^{N} M_i$ is the total number of observation units in the population

Design: One-Stage Cluster Sample

Why?

- Can be **cheaper** than a SRS
- A sampling frame of clusters may exist but a sampling frame of observation units does not.

Example 1: GPA

A student wants to estimate the average GPA in his dormitory. The dorm consists of 100 suites, each with four students. He chooses a SRS of 5 of these suites and records the GPA of each student living in the suite.

Example 2 - residents

Suppose you are interested in surveying the 11,482 adults who reside permanently in Northfield. You divide the town into 400 blocks and take a SRS of 5 blocks. You then visit each adult resident who lives on a selected block and record their annual income (in thousands of dollars) and whether or not they identify their political affiliation as Democratic.

Inclusion probabilities: One-Stage Cluster

What is the probability that unit j from cluster i is selected?

Sampling weights: One-Stage Cluster

What is the sampling weight for unit j from cluster i under a one-stage cluster design?

Estimation plan: One-Stage Cluster

- **One option!** Use an **unbiased** Horvitz-Thompson estimator to estimate the (overall) **population total**

$$\hat{t}_{HT} = \sum_{\text{sampled units}} w_{ij} y_{ij}$$

Population Total: One-Stage Cluster

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

Population Total: One-Stage Cluster

- **Parameter**: $t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} t_i$
- **Unbiased Estimator:**

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^{n} t_i = N\bar{t}$$

where $\bar{t}$ is the sample mean **total response** per cluster

Population Total: One-Stage Cluster

- **Parameter**: $t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} t_i$
- **Unbiased Estimator:**

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^{n} t_i = N\bar{t}$$

where $\bar{t}$ is the sample mean **total response** per cluster.

- **Standard error:**

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

where s_t is the sample standard deviation of cluster totals.

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \dfrac{t}{M_0}$
- **Assume that M_0 is known**
- **Unbiased Estimator:**

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

where $\bar{t}$ is the sample mean **total response** per cluster.

- **Standard error:**

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

Population Proportion: One-Stage Cluster

- **Parameter**: $p = \dfrac{t}{M_0}$
- Use formulas for mean where t_i counts the number of observation units in cluster i that are a "success"

Example 1 - GPA

$N = 100, n = 5, M_i = 4, M_0 = 400$

	Suite 1	Suite 2	Suite 3	Suite 4	Suite 5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
total	12.16	11.36	8.96	12.96	11.08

Estimate/SE for the mean GPA in the population.

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \frac{t}{M_0}$
- **What if M_0 is unknown!**

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \frac{t}{M_0}$
- **Assume that M_0 is unknown**
- **Biased Ratio Estimator:**

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} M_i}$$

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \frac{t}{M_0}$
- **Assume that M_0 is unknown**
- **Biased Ratio Estimator:**

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} M_i}$$

- **Standard error:** for large n:

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

Population Total: One-Stage Cluster

- **Parameter**: $t = M_0 \bar{y}_{\mathcal{U}}$
- **Assume that M_0 is known!!**
- **Biased Ratio Estimator:**

$$\hat{t}_r = M_0 \hat{\bar{y}}_r$$

- **Standard error:** for large n

$$SE(\hat{t}_r) \approx M_0 SE(\hat{\bar{y}}_r)$$

One-Stage Cluster estimation options:

- Unbiased vs. Biased:
  - Biased (ratio) options could be more precise than unbiased options when t_i and M_i are positively correlated
- Bias of ratio options:
  - need large n for bias to be small

Example 2 - residents

$N = 400, n = 5$

	Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**total**	s_t^2
# of Adults	10	15	18	22	17	82	19.3
Total Income	1100	1020	972	704	714	4510	33144
# Dems	8	5	7	15	3	38	20.8

Assume that $M_0 = 11,482$. Estimate/SE the proportion of adults who are Democrats.

Example 2 - residents

$N = 400, n = 5$

	Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**total**	s_t^2
# of Adults	10	15	18	22	17	82	19.3
Total Income	1100	1020	972	704	714	4510	33144
# Dems	8	5	7	15	3	38	20.8

Assume you don't know M_0. Estimate/SE the proportion of adults who are Democrats.