One-stage cluster sampling estimation

Week 5 (5.1, 5.2.1, 5.2.3)

Stat 260, St. Clair

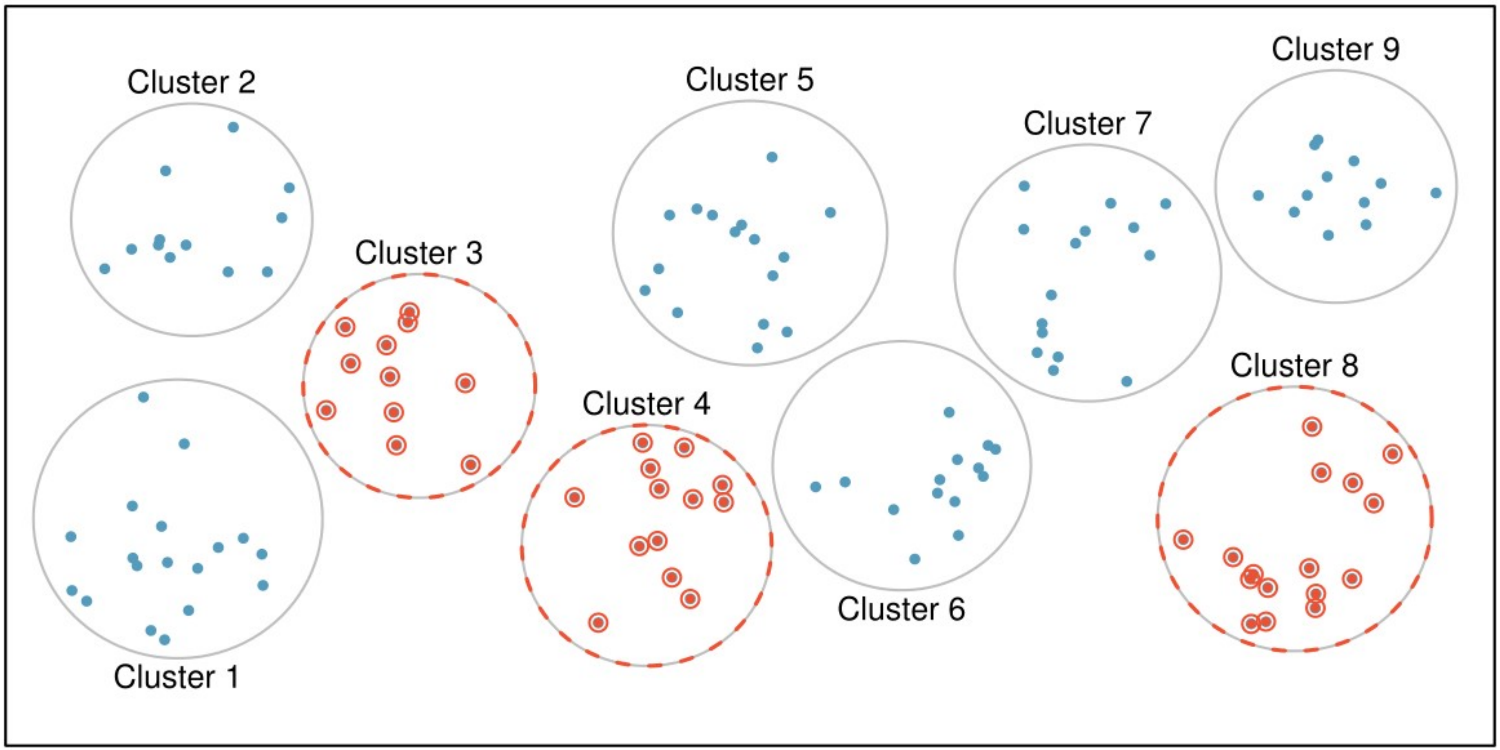
Design: One-Stage Cluster Sample

elements

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

measurement

$N = 9$

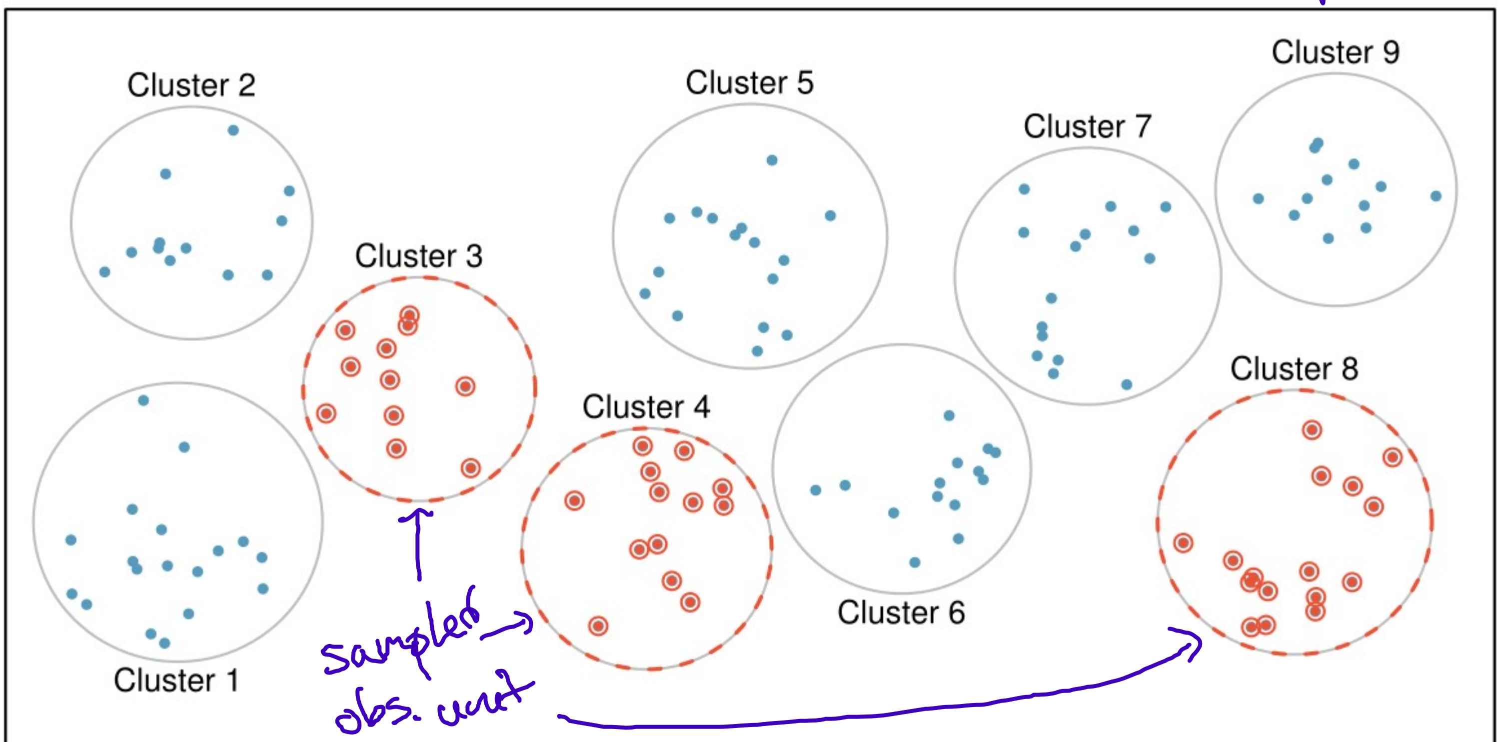


https://spot.pcc.edu/~evega/section-4.html

Design: One-Stage Cluster Sample

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

(census) $N = 9$



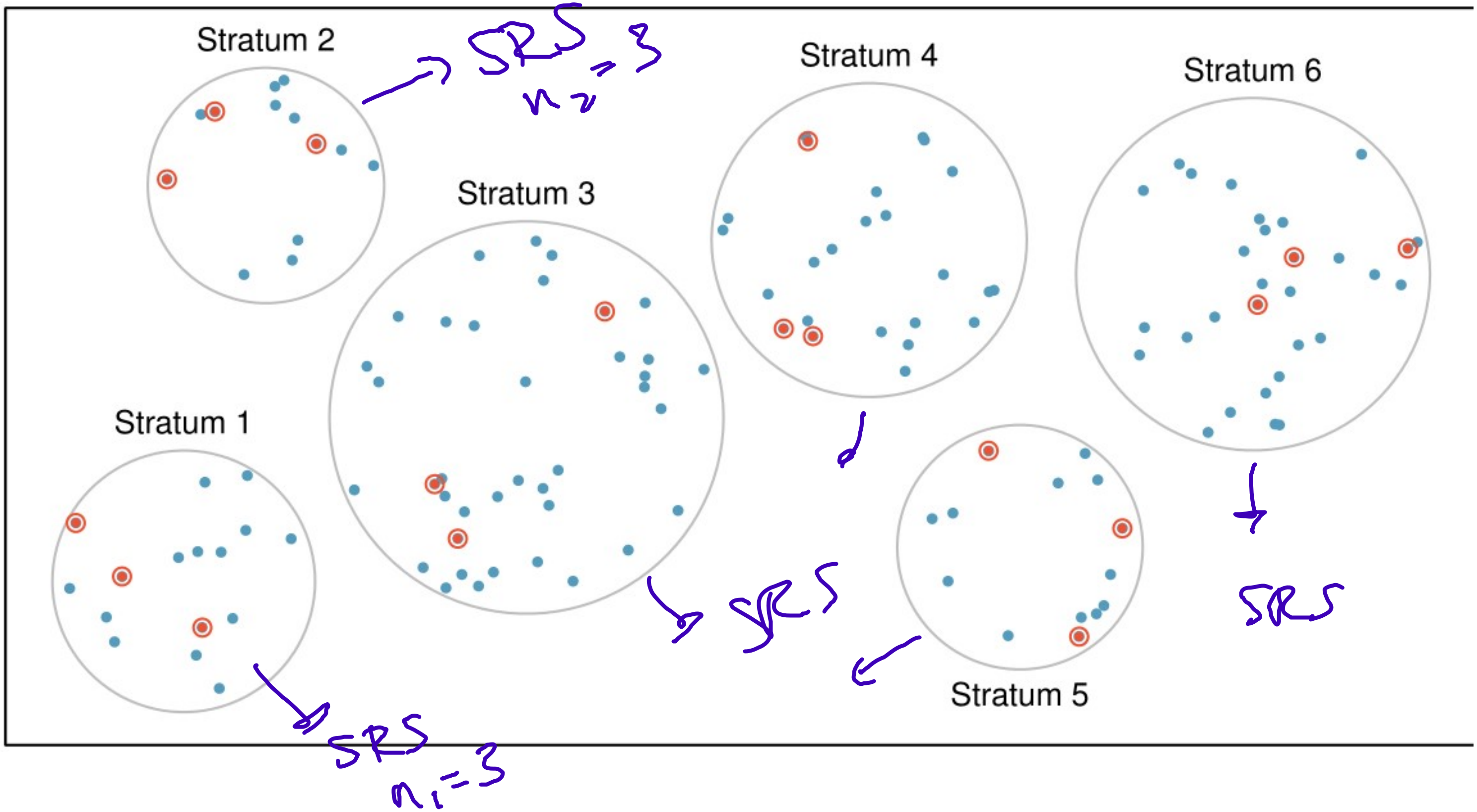
$n = 3$

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

Design: Cluster vs. Stratified sampling

Take a SRS **within** each strata

$H = 6$



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

Design: One-Stage Cluster Sample

- **Primary Sampling Units (PSU):** clusters (1)
- **Secondary Sampling Units (SSU):** observation units (2)
    - 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.

Cluster: suite

Observation unit: students

$N = 100$

$n = 5$

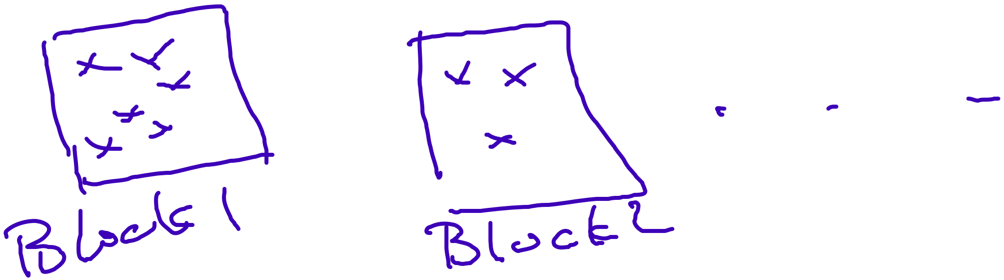
$M_i = 4$ for all i

$M_0 = \sum_{i=1}^{100} M_i = 100 \times 4 = 400$

y_{ij} = GPA of student j in suite i.

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.



cluster = Block
obs. unit = adult resident

$N = 400$

$n = 5$

M_i = # resid. in block i

$M_0 = 11{,}482$

Inclusion probabilities: One-Stage Cluster

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

given

$$\pi_{ij} = P(\text{unit } j \text{ from cluster } i \text{ selected})$$

$$= P(\text{cluster } i \text{ selected}) \times P(\text{unit } j \text{ selected} \mid \text{cluster } i \text{ selected})$$

$$= \boxed{\frac{n}{N} \times 1}$$

SRS of size n from N clusters

$\frac{n}{N}$

one-stage

all units selected

Sampling weights: One-Stage Cluster

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

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{1}{n/N} = \frac{N}{n}$$

at the obs. unit level:

$\frac{N}{n}$ = # of observation units in the pop. that unit j represents

Estimation plan: One-Stage Cluster

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

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

$$\hat{t} = \sum_{i=1}^{n} \sum_{j=1}^{M_i} \left(\frac{N}{n}\right) y_{ij} = \left(\frac{N}{n}\right) \sum_{i=1}^{n} t_i = N \bar{t}$$

i → clusters

j → unit

$\bar{t}$ = mean of cluster totals (sample)

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{cluster total}$$

$\frac{N}{n}$ = # of clusters represent by cluster i (t_i)

Population Total: One-Stage Cluster

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

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.

CI: t-dist. n-1 d.f.

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \dfrac{t}{M_0}$ $\Rightarrow$ mean response per obs. unit
- **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"

$$t_i = \sum_{i=1}^{M_i} y_{ij} = \text{\# of successes in cluster } i.$$

binary 1/0

Example 1 - GPA

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

→ clusters

	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

obs. units

→ t_i

Estimate/SE for the mean GPA in the population.

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i=1}^{n} t_i = \frac{100}{5}(12.16 + 11.36 + \dots + 11.08) = 1130.4$$

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}}{M_0} = \frac{1130.4}{400} = \boxed{2.83}$$

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

$$= \frac{100\sqrt{\left(1-\frac{5}{100}\right)\frac{s_t^2}{5}}}{400} = \boxed{.164}$$

$$s_t^2 = \frac{1}{n-1}\sum(t_i - \bar{t})^2 \qquad \bar{t} = \frac{12.16 + 11.36 + \cdots}{5} = 11.304$$

$$= \frac{1}{5-1}\left[(12.16 - 11.304)^2 + (11.36 - 11.304)^2 + \cdots\right]$$

$$= 2.256$$

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \dfrac{t}{M_0}$

- **What if M_0 is unknown!**

est. from our sample → SRS of clusters

M_i

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

$M_1, \ldots, M_n$ → SRS

$$\hat{M}_0 = N\,\overline{M} = \frac{N}{n}\sum_{i=1}^{n} M_i$$

SRS

sample mean cluster size

Population Mean: One-Stage Cluster

$$\frac{\tilde{t}}{\hat{M}_0} = \frac{\cancel{\frac{N}{n}} \sum t_i}{\cancel{\frac{N}{n}} \sum M_i}$$

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

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} M_i} = \frac{\text{total respone sampled}}{\text{total \# units sample}}$$

Ratio estimate

SE

Population Mean: One-Stage Cluster

- **Parameter**: $\bar{y}_{\mathcal{U}} = \dfrac{t}{M_0}$
- **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:

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

* Ratio SE

$\hat{B} = \dfrac{\Sigma y \longrightarrow t_i}{\Sigma x \longrightarrow M_i}$

$e_i = t_i - \hat{\bar{y}}_r M_i$

Population Total: One-Stage Cluster

- **Parameter**: $t = M_0\bar{y}_{\mathcal{U}}$ $= \sum^{N} \sum^{M_i} y_{ij}$
- **Assume that** $\boxed{M_0 \textbf{ is known!!}}$
- **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 $= t_1$	5 $= t_2$	7	15	3	38	20.8

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

unbiased est. of proportion:

total # dem.: $\hat{t}_{unb} = \frac{N}{n}\sum t_i = \frac{400}{5}(38) = 400(7.6) = 3040$

$\hat{p}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{3040}{11,482} = .265$

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

$$= \frac{400\sqrt{\left(1-\frac{5}{400}\right)\frac{20.8}{5}}}{11482} = .071$$

26.5%, $SE \approx 7\%$

unbiased (M_0 known)

Example 2 - residents

$N = 400, n = 5$

	Block 1	Block 2	Block 3	Block 4	Block 5	total	s_t^2
M_i # of Adults	10 $= M_1$	15 $= M_2$	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.

Ratio biased

$$\hat{p}_r = \frac{\sum t_i}{\sum M_i} = \frac{38}{82} = .463$$

$$SE(\hat{p}_r) = \sqrt{\left(1 - \frac{5}{400}\right)\frac{1}{5(\bar{M})^2} \times s_e^2} = .108$$

$$\bar{M} = \frac{\Sigma M_i}{5} = \frac{82}{5} = 16.4$$

$$s_e^2 = \frac{1}{5-1}\Sigma\left(t_i - \hat{p}_r M_i\right)^2$$

$$= \frac{1}{4}\left(\left(8 - \left(\frac{38}{82}\right)(10)\right)^2 + \cdots\right)$$

$$= 15.955$$

Ratio/Biased: 46 %, SE $\approx$ 11%