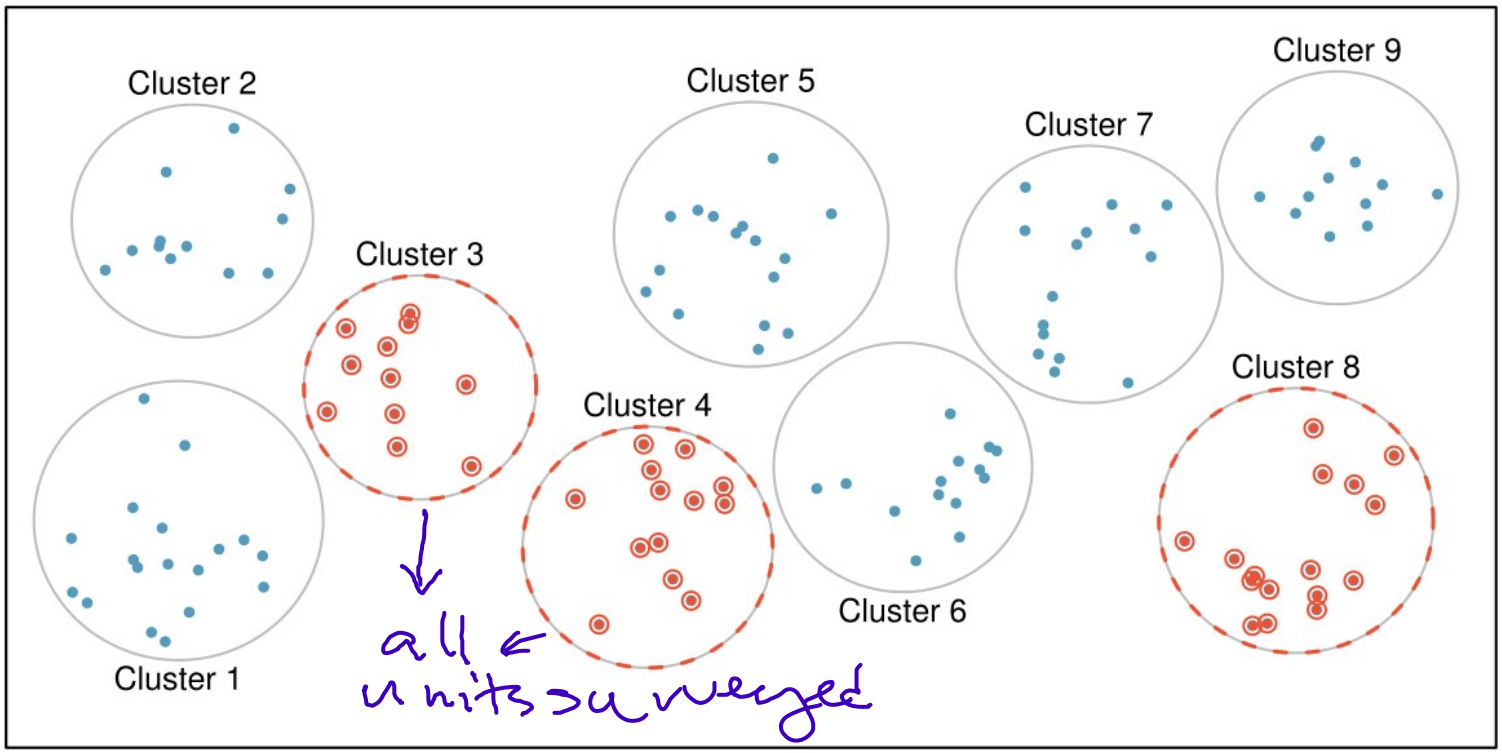
Two-stage cluster sampling estimation

Week 7 (5.3)

Stat 260, St. Clair

Design: One-Stage Cluster Sample

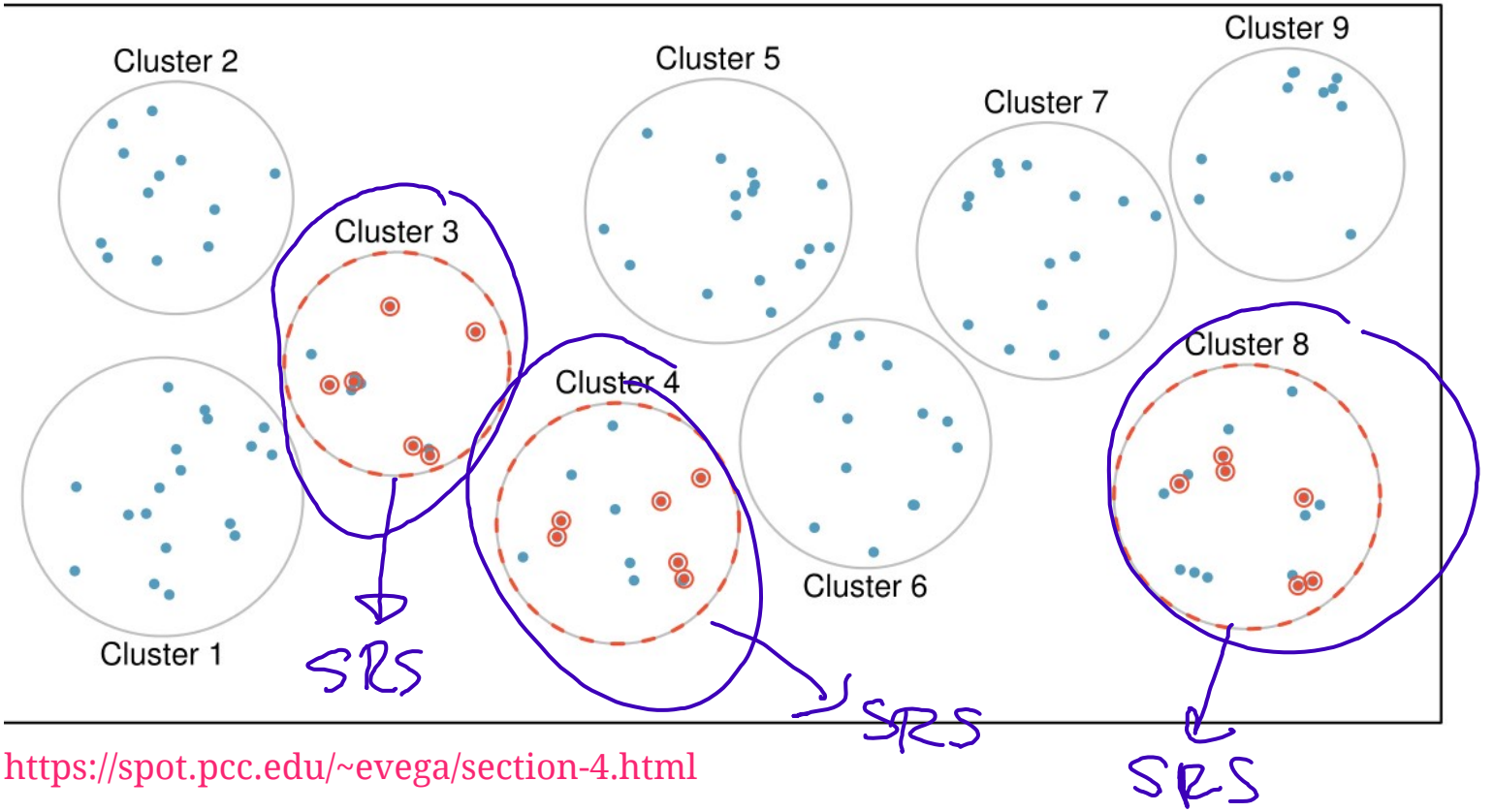
Defined: We take a SRS of n **clusters** and survey **every observation unit** in selected clusters.



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

Design: Two-Stage Cluster Sample

Defined: We take a **SRS of clusters** and take a **SRS of observation units** within each selected cluster.



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

Design: Two-Stage Cluster Sample

- **Primary Sampling Units (PSU):** clusters
  - N clusters in the population
  - n number of clusters sampled
- **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
  - $m_i \leq M_i$ is the sampled observation units in cluster i
  - $M_0 = \sum_{i=1}^{N} M_i$ is the total number of observation units in the population

Example: California API scores

1

- A SRS of 40 school districts was selected from the 757 districts in the state. `dname` gives district name and `dnum` is a numerical ID for each district.
- Data from a SRS of schools within each selected district was collected. `sname` gives a 40 character name and `snum` gives a numerical ID for each school.

1 cluster = district

$N = 757$ $n = 40$

2 Obs. units = schools

m_i = # of schools sampled in district i

M_i = total # schools in district i

EDA → $1 \le m_i \le 5$

Inclusion probabilities: Two-Stage Cluster

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

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

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

① SRS size n from N ② SRS of m_i from M_i

$$= \underbrace{\frac{n}{N} \times \frac{m_i}{M_i}}_{① \quad ②}$$

Sampling weights: Two-Stage Cluster

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

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{N}{n} \times \frac{M_i}{m_i}$$

$\frac{N}{n}$: # of clusters represented by one sampled cluster.

$\frac{M_i}{m_i}$: # of units in cluster i represented by one sampled unit

Estimation plan: Two-Stage Cluster

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

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

$$\hat{t}_{HT} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{N}{n} \frac{M_i}{m_i} y_{ij} = \sum_{i=1}^{n} \frac{N}{n} M_i \bar{y}_i$$

$\frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} = \bar{y}_i$ = sample avg. in cluster i

$\hat{t}_i = M_i \bar{y}_i$ = SRS est. of cluster i total t_i

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{N}{n} \hat{t}_i$$

one-stage est.

$$\hat{t}_{HT} = \sum \frac{N}{n} t_i$$

Population Total: Two-Stage Cluster

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

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

  where $\hat{t}_i$ is the *estimated* total response in cluster i.

- **Standard error:**

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

  where s_t is the sample standard deviation of *estimated* cluster totals and s_i is the sample SD within cluster i.

Population Mean: Two-Stage Cluster

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

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

- **Standard error:**

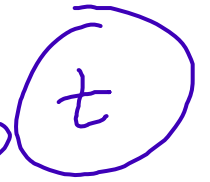
$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{M_0}$$

Population Proportion: Two-Stage Cluster

- **Parameter**: $p = \dfrac{t}{M_0}$
- Use formulas for mean where
  - $\bar{y}_i = \hat{p}_i$ is cluster sample proportion
  - $\hat{t}_i$ estimates the number of observation units in cluster i that are a "success"
  - $s_i^2 = \dfrac{m_i}{m_i - 1}\hat{p}_i(1 - \hat{p}_i)$

(SRS)

Example: California API scores



Estimate the number of schools that have over 50% of students who are eligible for a subsidized meal.

```
> schools$meals_level <- ifelse(schools$meals > 50, "above 50%", "50%
> glimpse(schools)
Rows: 126
Columns: 11
$ sname          <chr> "Alta-Dutch Flat Elementary", "Tenaya Elementar
$ snum           <int> 3269, 5979, 4958, 4957, 4956, 4915, 2548, 2550,
$ dname          <chr> "Alta-Dutch Flat Elem", "Big Oak Flat-Grvlnd Ur
$ dnum           <int> 15, 63, 83, 83, 83, 117, 132, 132, 132, 152, 15
$ api00          <int> 821, 773, 600, 740, 716, 811, 472, 520, 568, 59
$ growth         <int> 36, 55, -32, 0, 5, 32, 40, 26, -21, 6, -10, 29,
$ meals          <int> 27, 43, 33, 11, 5, 25, 78, 76, 68, 42, 63, 54,
$ ell            <int> 0, 0, 5, 4, 2, 5, 38, 34, 34, 23, 42, 24, 3, 6,
$ enroll         <int> 152, 312, 173, 201, 147, 234, 184, 512, 543, 33
$ district_size  <int> 1, 1, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4,
$ meals_level    <chr> "50% or below", "50% or below", "50% or below",
```

Example: California API scores

$\hat{t}_{unb} = \sum \frac{N}{n} M_i \bar{y}_i$

$= \sum \frac{N}{n} M_i \hat{p}_i$

$\hat{t}_i =$ est. total # schools above 50% meals.

```
> schools_by_district <- schools %>%
+   group_by(dnum) %>%  # group by cluster (district number)
+   summarize(p_hat = mean(meals_level == "above 50%"),
+             m_i = n(), # sample size per cluster
+             M_i = first(district_size), # pop size per cluster
+             t_hat_i = M_i*p_hat) %>%
+   arrange(desc(M_i))   # arrange by big to small clusters
> kable(schools_by_district, digits = 2)
```

$\hat{p}_i$ m_i M_i $\hat{t}_i$

i	dnum	p_hat	m_i	M_i	t_hat_i
1	620	1.00	5	72	72.0 $= \hat{t}_1$
2	570	0.80	5	36	28.8 $= \hat{t}_2$
	575	0.00	5	28	0.0
	638	0.40	5	14	5.6
	200	1.00	5	11	11.0
	731	0.20	5	9	1.8
	596	0.00	5	7	0.0

Example: California API scores

Estimate the number of schools that have over 50% of students who are eligible for a subsidized meal.

```
> N <- 757  # number of clusters in pop
> n <- 40   # number of clusters sampled
> schools_by_district %>%
+   summarize(t_unb = (N/n)*sum(t_hat_i))
# A tibble: 1 x 1
  t_unb
  <dbl>
1 2729.
```

$$= \sum_{i=1}^{40} \frac{757}{40} \hat{t}_i$$

Example: California API scores

* schools -> "raw" data

But, we should be using the survey package instead:

```
> schools$N <- 757  # N
> schools$n <- 40   # n
> schools <- schools %>%
+     group_by(dnum) %>% # group by cluster (district number)
+     mutate(m_i = n())  #  m_i = sample size per cluster
> summary(schools$m_i)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       3       5       4       5       5
> schools$wts <- (757/40)*schools$district_size/schools$m_i
> summary(schools$wts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.93   18.93   18.93   40.70   26.50  272.52
```

N/n M_i/m_i

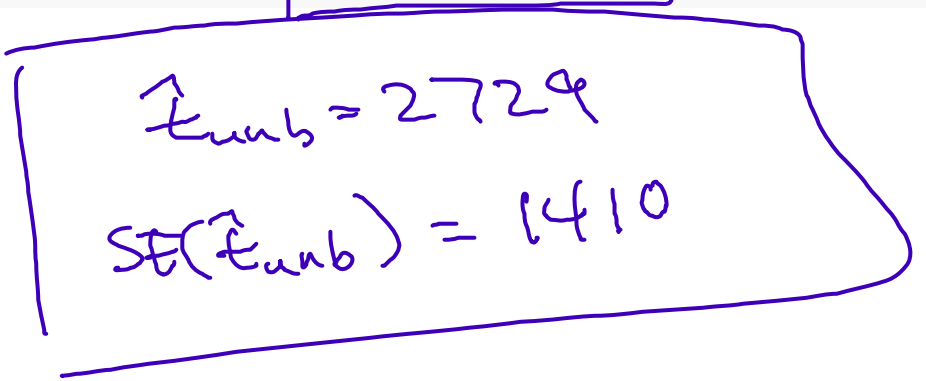
M_i

mutate preserves size of schools data -> rows = schools

Example: California API scores

But, we should be using the `survey` package instead:

```
> schools_design <- svydesign(id= ~dnum + snum,
+                             fpc= ~N + district_size,
+                             weights = ~wts,
+                             data=schools)
> svytotal(~meals_level, schools_design, deff = TRUE)
                            total      SE    DEff
meals_level50% or below 2399.69  551.19  5.9451
meals_levelabove 50%    2728.99 1410.37 38.9246
```



Population Mean: Two-Stage Cluster

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

Population Mean: Two-Stage Cluster

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

$$\hat{\bar{y}}_r = \frac{\sum_{i=1}^{n} \hat{t}_i}{\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i} \quad \approx \quad \frac{\sum\sum \frac{N}{n} \frac{M_i}{m_i} y_{ij}}{\sum\sum \frac{N}{n} \frac{m_i}{m_i}}$$

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

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

(1) (2)

Population Total: Two-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

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

Example: California API scores

Estimate the *proportion* of schools that have over 50% of students who are eligible for a subsidized meal.

$$p = \frac{\#\text{ of schools meals} > 50\%}{\text{total \# schools}}$$

??

Example: California API scores

$$\hat{t}_r = \frac{\sum \hat{t}_i}{\sum M_i} = \frac{72 + 28.8 + 0 + \ldots}{72 + 36 + 28 + \ldots}$$

```
> kable(schools_by_district, digits = 2)
```

dnum	p_hat	m_i	M_i	t_hat_i
620	1.00	5	72	72.0
570	0.80	5	36	28.8
575	0.00	5	28	0.0
638	0.40	5	14	5.6
200	1.00	5	11	11.0
731	0.20	5	9	1.8
596	0.00	5	7	0.0
639	0.00	5	7	0.0
781	0.00	5	6	0.0
295	0.00	5	5	0.0
403	0.00	5	5	0.0
480	0.40	5	5	2.0

Example: California API scores

Estimate the *proportion* of schools that have over 50% of students who are eligible for a subsidized meal.

```
> N <- 757  # number of clusters in pop
> n <- 40   # number of clusters sampled
> schools_by_district %>%
+   summarize(p_hat_r = sum(t_hat_i)/sum(M_i))
# A tibble: 1 x 1
  p_hat_r
    <dbl>
1   0.532
```

$= \hat{p}_r$

Example: California API scores

But, we should be using the `survey` package instead:

```
> svymean(~meals_level, schools_design, deff = TRUE)
                                mean      SE DEff
meals_level50% or below 0.46790 0.14486 10.8
meals_levelabove 50%    0.53210 0.14486 10.8
```

$\hat{p}_r = .532$

$SE(\hat{p}_r) = .145$