Two-stage cluster sampling estimation: variance

Week 7 (5.3)

Stat 260, St. Clair

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

- **Standard error:**

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

Variance: between + within cluster variation

1. Between cluster variation: $N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}$

- just the **one-stage** variance

$n \approx N$, then (1) small & $SE(\hat{t}_{unb})$ similar to a stratified SE.

2. Within cluster variation: $\frac{N}{n}\sum_{i=1}^{n}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$

- added because we don't sample *all* units within a cluster

$m_i \approx M_i \Rightarrow$ then (2) small & $SE(\hat{t}_{unb})$ similar to one-stage SE

- Same explanation of the *ratio* estimator $SE(\hat{\bar{y}}_r)$

Comparing cluster sampling and SRS

- **One-stage** *less* precise than a SRS when clusters are *homogeneous*
  - Same holds true for **two-stage**
  - Generally, $SE_{two-stage} \geq SE_{one-stage} > SE_{SRS}$

Two-stage cluster sampling: Assessing homogeneity

- Visually: plot cluster id vs y_{ij} (boxplot, scatterplot)
- Adjusted R-squared:

$$\hat{R}_a^2 = 1 - \frac{\widehat{MSW}}{\hat{S}^2}$$

- MSW is estimated from **sample ANOVA** values msw

M m

- **Cluster/sample sizes equal:** S^2 is estimated from **sample ANOVA** values msb and msw:

$$\hat{S}^2 = \frac{\frac{M}{m}(N-1)msb + \left(\frac{m-1}{m}NM + \frac{M}{m} - 1\right)msw}{NM - 1}$$

- **Cluster/sample sizes NOT equal:** best "by hand" (but biased) approximation is the sample variance of all responses s^2

Example: California API scores by district

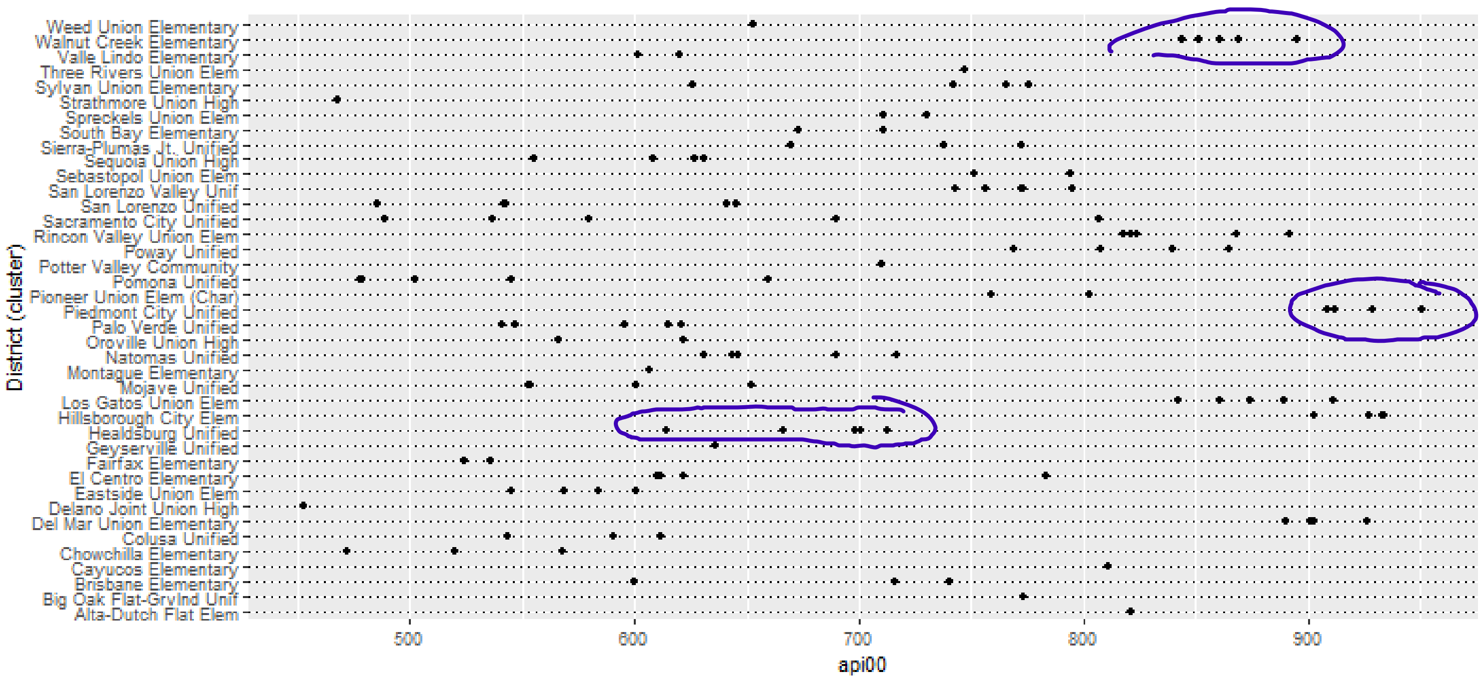
```
> schools_by_district <- schools %>%
+   group_by(dnum) %>%  # group by cluster (district number)
+   summarize( s_i = sd(api00),     # SD by cluster
+              m_i = n(), # sample size per cluster
+              M_i = first(district_size) ,
+              samp_frac = m_i/M_i)
> summary(schools_by_district)
      dnum             s_i                m_i              M_i
 Min.   : 15.0   Min.   :  8.485   Min.   :1.00   Min.   : 1.000
 1st Qu.:221.0   1st Qu.: 20.769   1st Qu.:1.75   1st Qu.: 1.750
 Median :541.5   Median : 34.894   Median :3.00   Median : 3.000
 Mean   :463.7   Mean   : 39.464   Mean   :3.15   Mean   : 6.775
 3rd Qu.:675.2   3rd Qu.: 47.753   3rd Qu.:5.00   3rd Qu.: 5.000
 Max.   :795.0   Max.   :127.982   Max.   :5.00   Max.   :72.000
                 NA's   :10
   samp_frac
 Min.   :0.06944
 1st Qu.:1.00000
 Median :1.00000
 Mean   :0.87540
 3rd Qu.:1.00000
 Max.   :1.00000
```

$\Rightarrow m_i = M_i$

Example: California API scores by district

similar api scores for schools in same district ⇒ homogeneity

```
> ggplot(schools, aes(x=dname, y = api00)) +
+   geom_point() +
+   geom_vline(aes(xintercept=dname), linetype= "dotted") +
+   coord_flip() +
+   labs(x="District (cluster)")
```



Example: California API scores by district

- Cluster (district) sizes in the population are *not* similar.
  - quick approximation gives $R_a^2 \approx 0.86$ → high homogeneity

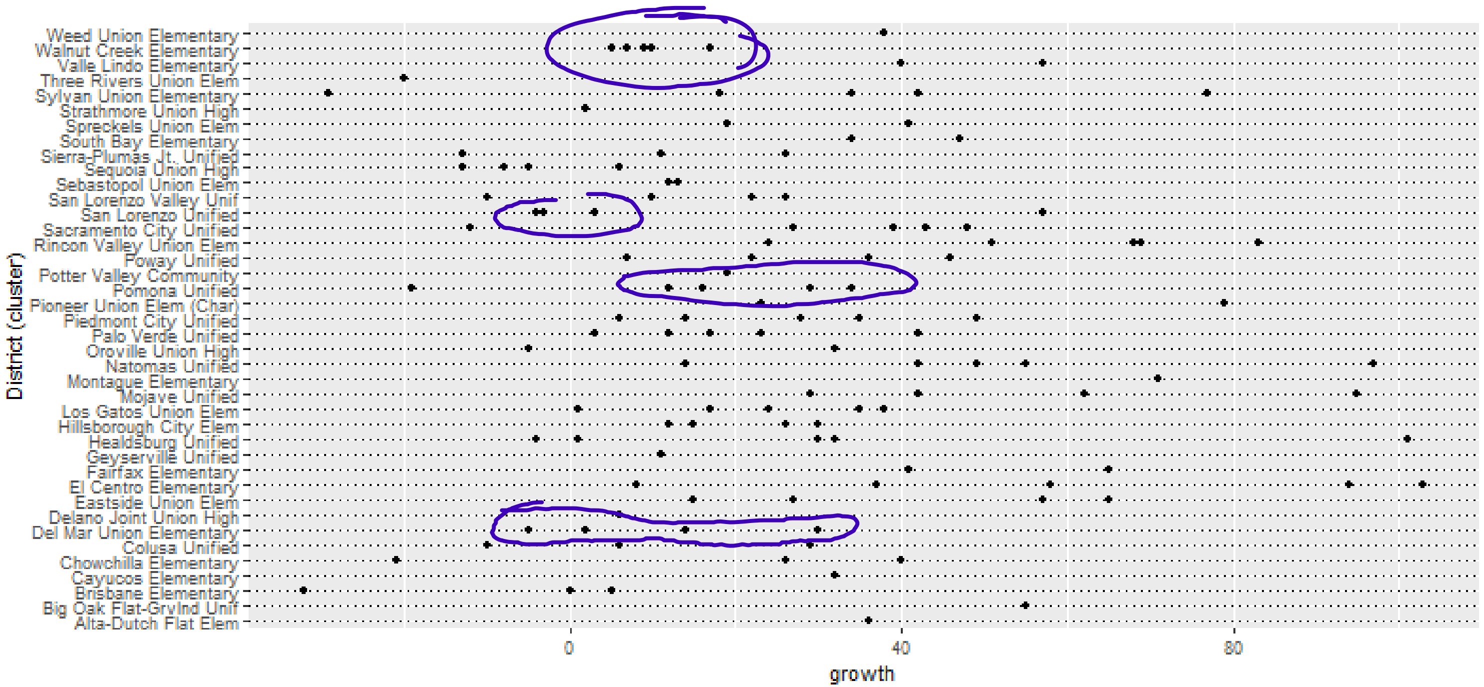
```
> api_lm <- lm(api00 ~ dname, data=schools)
> anova(api_lm)
Analysis of Variance Table

Response: api00
          Df  Sum Sq Mean Sq F value    Pr(>F)
dname     39 2028811   52021  20.267 < 2.2e-16 ***
Residuals 86  220740    2567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 1 - 2567/var(schools$api00)    # r^2_a =  1=msw/s^2
[1] 0.8573605
> summary(api_lm)$adj.r.squared  # r^2_a
[1] 0.8573744
```

within → Residuals; MSW

Example: California growth by district

```
> ggplot(schools, aes(x=dname, y = growth)) +
+   geom_point() +
+   geom_vline(aes(xintercept=dname), linetype= "dotted") +
+   coord_flip() +
+   labs(x="District (cluster)")
```



Example: California growth by district

- quick approximation gives $R_a^2 \approx 0.24$ $\Rightarrow$ less homogeneity than api variable

```
> growth_lm <- lm(growth ~ dname, data=schools)
> summary(growth_lm)$adj.r.squared
[1] 0.2448077
```

Example: California growth by district

- `growth` is more heterogeneous by district so its SE is more similar to a SRS than `api00`

```
> # design from estimation slides
> svymean(~api00 + growth, schools_design, deff = TRUE)
          mean      SE   DEff
api00  670.812  30.099 6.2505
growth  25.778   2.842 1.5794
```

→ closer to 1 ⇒

SE_{growth} 2-stage ≈ SE if SRS used

Variance: between + within cluster variation

(true pop)

$$Var(\hat{t}_{unb}) = \underbrace{N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}}_{A} + \underbrace{\frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}}_{B}$$

Sketch of Proof Uses Appendix Section A.4 Rule 5:

$$Var(Y) = Var_{\underset{x}{\cancel{y}}}(E_{\underset{y}{\cancel{x}}}(Y \mid X)) + E_{\underset{x}{\cancel{y}}}(Var_{\underset{y}{\cancel{x}}}(Y \mid X))$$

$$Var(\hat{t}_{unb}) = \underbrace{Var_1\left(E_2(\hat{t} \mid \text{stage } 1)\right)}_{A} + \underbrace{E_1\left(Var_2(\hat{t} \mid \text{stage } 1)\right)}_{B}$$

A. between cluster variation

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

Fixed clusters (stage 1): $i \in \mathcal{S}$ are fixed

$$E_2(\hat{t} \mid \text{stage 1 fixed})$$

$$E_2[\hat{t} \mid \text{stg.}1] = E\left(\frac{N}{n} \sum_{i \in S} \hat{t}_i \,\middle|\, \text{stg.}1\right) = \frac{N}{n} \sum_{i \in S} E[\hat{t}_i \mid \text{stg.}1]$$

$\hat{t}_i = M_i \bar{y}_i =$ SRS est. of $t_i \Rightarrow E[\hat{t}_i \mid \text{stg.}1] = t_i$

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

A. between cluster variation

How does E_2 vary across stage 1 samples?

$$Var_1(E_2(\hat{t} \mid \text{stage 1 fixed}))$$

A

$$Var_1\left(\frac{N}{n}\sum_{i\in S} t_i\right) = Var_1(N\bar{t}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}$$

SRS of clusters $\frac{N}{n}\sum_{i\in S} t_i = N\bar{t}$ = SRS est. of $t = \sum_{i=1}^{N} t_i$

$\Rightarrow$ SRS SE^2

B. within cluster variation

Fixed clusters (stage 1): $i \in \mathcal{S}$ are fixed

$$Var_2(\hat{t} \mid \text{stage 1 fixed})$$

$$Var_2\left(\frac{N}{n}\sum_{i\in S}\hat{t}_i \,\middle|\, \text{stg. 1}\right) = \left(\frac{N}{n}\right)^2 \sum_{i\in S} Var(\hat{t}_i)$$

$\hat{t}_i$: SRS est. of t_i $\Rightarrow$ SRS SE^2 ch. 2

$$= \left(\frac{N}{n}\right)^2 \sum_{i\in S} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

B. within cluster variation

What is the expected V_2 across stage 1 samples?

$$E_1(Var_2(\hat{t} \mid \text{stage 1 fixed}))$$

$$E_1\left[\left(\frac{N}{n}\right)^2 \sum_{i\in S} \underbrace{M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}}_{u_i}\right] = \left(\frac{N}{n}\right) E\left[\overbrace{\frac{N}{n}\sum_{i\in S} u_i}^{\bar{u}}\right]$$

$$= \frac{N}{n} E[N\bar{u}] = \frac{N}{n}\sum_{i=1}^{N} u_i = \boxed{\frac{N}{n}\sum_{i=1}^{N} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}}$$

$N\bar{u}$ = SRS est. of total $\sum_{i=1}^{N} u_i$