Survey package: One stage cluster sampling

Week 6

Stat 260, St. Clair

One-stage cluster sampling estimation



How you analyze one-stage cluster sampling data depends on the data format

SSU-level data: each row is a different measurement unit (SSU)

- svytotal will give you unbiased total estimates \hat{t}_{unb}
 - $\circ~$ for **mean**: divide by M_0 , if known, to get $\hat{ar{y}}_{unb}=\hat{t}_{\,unb}/M_0$, SE and CI
- svymean will give you ratio (biased) mean estimates $\hat{ar{y}}_r$
 - $\circ~$ for **total**: multiply by M_0 , if known, to get $\hat{t}_{\,r} = M_0 \hat{ar{y}}_{\,r}$, SE and CI

svymean

<u>Zwiyi</u>

Zwi -> Mo

(#students/class)

- class: class identifier (cluster)
- Mi: class size (cluster size)

• score: student (PSU) level test score

N = number of classes in the population (187)

```
> algebra$N <- 187 # number of clusters in pop.
> nrow(algebra) # number of sampled SSU = NOT "n"

[1] 299

H Shdents (SSN) sampled
```

```
n = number of sampled classes (12)
```

```
> (algebra$n <- n_distinct(algebra$class) ) # n
[1] 12
> algebra$wts <- algebra$N/algebra$n # weights N/n</pre>
```

```
> alg_design<- svydesign(id = ~class,</pre>
                        fpc= ~N,
+
                        weights= ~wts,
+
                        data=algebra)
> summary(alg_design)
1 - level Cluster Sampling design
With (12) clusters.
svydesign(id = ~class, fpc = ~N, weights = ~wts, data = algebra)
Probabilities:
  Min. 1st Qu. Median Mean 3rd Qu. Max.
0.06417 0.06417 0.06417 0.06417 0.06417 0.06417
Population size (PSUs): 187
Data variables:
[1] "class" "Mi" "score" "N" "n" "wts"
```

$$\frac{\hat{q}_r}{\sum_{i=1}^{N} \frac{\sum_{j=1}^{N} \binom{N}{j} \gamma_{ij}}{\sum_{j=1}^{N} \binom{N}{j}} = 62.6$$

$$SE(\hat{q}_r) = 1.5$$

$$\hat{t}_{unb} = \sum_{i,j} \left(\frac{N}{n} \right) y_{i,j} = 291,533$$

$$\left(est. \text{ total score in } 90P \right)$$

$$\hat{y}_{unb} = \frac{\hat{t}_{inb}}{M_0} = \frac{291,533}{M_0}$$

$$SF(\hat{y}_{unb}) = \frac{19,893}{M_0}$$

One-stage cluster sampling estimation

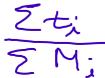
11 Summanized Lata

How you analyze one-stage cluster sampling data depends on the **data format**

PSU-level data: each row is a different cluster (PSU) total t_i and cluster size

```
> clus_design2 <- svydesign(id = ~1, # 1 = row = PSU</pre>
                              fpc = \sim N, # N = number of clusters
                              weights = \simwts, # N/n
                              data = clusterdata2)
> svyratio(~t, ~M, clus_design2) # ratio mean estimate
> svytotal(~t, clus_design2  # unbiased total estimate
                                                        = \sum \left(\frac{n}{n}\right) f
```

- svytotal will give you unbiased total estimates \hat{t}_{unb}
 - $\circ~$ for **mean**: divide by M_0 , if known, to get $\hat{ar{y}}_{unb}=\hat{t}_{\,unb}/M_0$, SE and CI
- ullet svyratio will give you ratio (biased) mean estimates $\hat{ar{y}}_r$
 - $\circ~$ for **total**: multiply by M_0 , if known, to get $\hat{t}_{\,r} = M_0 \hat{ar{y}}_{\,r}$, SE and CI



1/

N =	400, n	= 5	
		_	_

- Unother					1		
	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

cd. -1. -

Block (cluster) level data:

Estimate/SE the proportion of adults who are Democrats. Assume M_0 is unknown.

```
> svyratio(~dem_tots, ~block_size, block_design) # ratio estimate

Ratio estimator: svyratio.survey.design2(~dem_tots, ~block_size, block_size)

block_size

dem_tots 0.4634146 = pr = gr

SEs=

block_size

dem_tots 0.1082384 = SE(pr)
```

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

```
> t_unb <- svytotal(~dem_tots, block_design) # unbiased estimate</pre>
> t_unb
        total SE
dem_tots 3040 810.73
> coef(t_unb)/11482 # unbiased proportion
 dem_tots
0.2647622
> SE(t_unb)/11482 # SE
dem_tots
dem_tots 0.07060861
          3040 (4 Dems in pap.)
```