`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)

```
> clus_design1 <- svydesign(id = ~PSU,
+                           fpc = ~N,     # N = number of clusters
+                           weights = ~wts,    # N/n
+                           data = clusterdata1)
> svymean(~y, clus_design1)    # ratio mean estimate
> svytotal(~y, clus_design1)   # unbiased total estimate
```

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

Lohr Examples 5.6

```
> algebra <- read.csv("http://math.carleton.edu/kstclair/data/algebra
> library(dplyr)
> glimpse(algebra)
Rows: 299
Columns: 3
$ class <int> 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23,
$ Mi    <int> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
$ score <int> 57, 90, 56, 57, 46, 55, 62, 66, 78, 76, 57, 84, 27, 70,
```

numeric

- `class`: class identifier (cluster)
- `Mi`: class size (cluster size) (# students / class)
- `score`: student (PSU) level test score

y_{ij}

Lohr Examples 5.6

`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
```

→ # students (SSU) sampled

`n` = number of sampled classes (12)

dplyr

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

Lohr Examples 5.6


```
> alg_design<- svydesign(id = ~class,
+                        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"
```

Lohr Examples 5.6

```
> svymean(~score, alg_design)  # ratio estimate/SE of pop.mean score
        mean     SE
score 62.569 1.4916
> confint(svymean(~score, alg_design), df=degf(alg_design))
         2.5 %  97.5 %
score 59.28562 65.8515
> degf(alg_design)   # n - 1
[1] 11
```

$$\hat{\bar{y}}_r = \frac{\sum_i \sum_j \left(\frac{N}{n}\right) y_{ij}}{\sum_i \sum_j \left(\frac{N}{n}\right)} = 62.6$$

$$SE(\hat{\bar{y}}_r) = 1.5$$

Lohr Examples 5.6

```
> svytotal(~score, alg_design)  # unbiased estimate/SE of pop.total
       total    SE
score 291533 19893
```

$$\hat{t}_{unb} = \sum_i \sum_j \left(\frac{N}{n}\right) y_{ij} = 291,533$$

(est. total score in pop)

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

$$SE\left(\hat{\bar{y}}_{unb}\right) = \frac{19,893}{M_0}$$

??

One-stage cluster sampling estimation

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 M_i

```
> clus_design2 <- svydesign(id = ~1,      # 1 = row = PSU
+                           fpc = ~N,     # N = number of clusters
+                           weights = ~wts,   # N/n
+                           data = clusterdata2)
> svyratio(~t, ~M, clus_design2)  # ratio mean estimate
> svytotal(~t, clus_design2)    # unbiased total estimate
```

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

Week 5 example - residents

obs. units = people

$N = 400, n = 5$

cluster

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

Block (cluster) level data:

```
> block_data <- data.frame(
+   dem_tots = c( 8, 5, 7, 15, 3),
+   block_size = c(10, 15, 18, 22, 17)
+ )
> block_data
  dem_tots block_size
1        8         10
2        5         15
3        7         18
4       15         22
5        3         17
```

block: t_i dem_tots, M_i block_size

summarized by cluster (block)

Week 5 example - residents

```
> block_data$N <- 400
> block_data$n <- 5
> block_data$wts <- 400/5
> block_design <- svydesign(id = ~1,
+                           fpc = ~N,     # N = number of blocks
+                           weights = ~wts,    # N/n
+                           data = block_data)
```

summarized

Week 5 example - residents

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

t_i M_i

```
> svyratio(~dem_tots, ~block_size, block_design) # ratio estimate
Ratio estimator: svyratio.survey.design2(~dem_tots, ~block_size, bloc
Ratios=
          block_size
dem_tots   0.4634146
SEs=
          block_size
dem_tots   0.1082384
```

$= \hat{p}_r = \hat{\bar{y}}_r$

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

Week 5 example - residents

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
> 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
```

$\hat{p}_{unb}$

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

$\hat{t}_{unb} = 3040$ (# Dems in pop)

$\hat{p}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{3040}{11482} = .265$