One-stage cluster sampling simulation

Math 255 - St. Clair

1. The population

The values in `Sim_Cluster_Pops.csv` represent a simulated population with response y and three possible clustering variables `cluster1`, `cluster2` and `cluster3`.

```
> pop <- read.csv("http://math.carleton.edu/kstclair/data/Sim_Cluster_Pops.csv")
> str(pop)
'data.frame':	500 obs. of  5 variables:
 $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ y       : num  0.0125 0.0678 0.1084 0.1757 0.2034 ...
 $ cluster1: int  54 10 35 41 96 29 26 48 80 76 ...
 $ cluster2: int  1 2 3 4 5 6 7 8 9 10 ...
 $ cluster3: int  1 1 1 1 1 2 2 2 2 2 ...
```

The population has the following characteristics:

- $N = 100$ clusters for each clustering variable option

```
> library(tidyverse)
> n_distinct(pop$cluster1)
[1] 100
> n_distinct(pop$cluster2)
[1] 100
> n_distinct(pop$cluster3)
[1] 100
```

- $M_i = M = 5$ elements per cluster for each clustering variable option

```
> pop %>% group_by(cluster1) %>% count() %>% ungroup() %>% summary()
    cluster1            n    
 Min.   :  1.00   Min.   :5  
 1st Qu.: 25.75   1st Qu.:5  
 Median : 50.50   Median :5  
 Mean   : 50.50   Mean   :5  
 3rd Qu.: 75.25   3rd Qu.:5  
 Max.   :100.00   Max.   :5  
> pop %>% group_by(cluster2) %>% count() %>% ungroup() %>% summary()
    cluster2            n    
 Min.   :  1.00   Min.   :5  
 1st Qu.: 25.75   1st Qu.:5  
 Median : 50.50   Median :5  
 Mean   : 50.50   Mean   :5  
 3rd Qu.: 75.25   3rd Qu.:5  
 Max.   :100.00   Max.   :5  
> pop %>% group_by(cluster3) %>% count() %>% ungroup() %>% summary()
    cluster3            n    
 Min.   :  1.00   Min.   :5  
 1st Qu.: 25.75   1st Qu.:5  
 Median : 50.50   Median :5  
 Mean   : 50.50   Mean   :5  
 3rd Qu.: 75.25   3rd Qu.:5  
```

```
 Max.   :100.00   Max.   :5
```

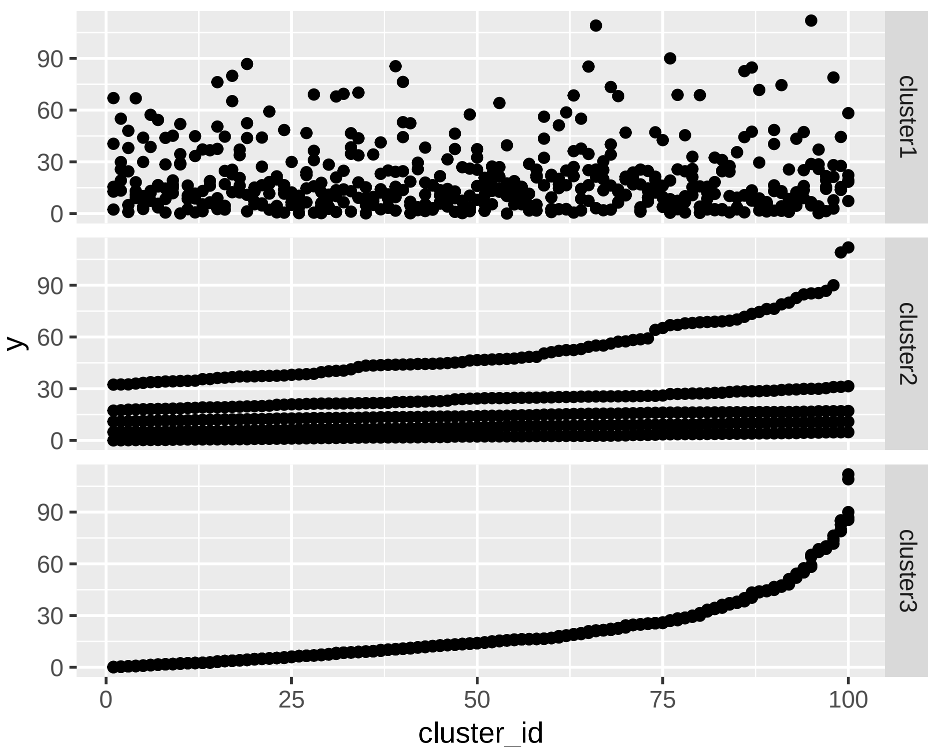
- $M_0 = NM = 500$ elements in the population

2. Simulation goals

1. For a given response and clustering variable, compare precision of a one stage cluster sample of $n = 5$ clusters (with $nM = 25$ elements) to a SRS of $n = 25$ elements.
2. How does 1 depend on the clustering variable?

The following code chunk plots the reponse y by cluster ID for the three cluster variable options.

```
> pop_long <- pop %>% select(-X) %>% gather(key = cluster_type,
+     value = cluster_id, cluster1:cluster3)
> str(pop_long)
'data.frame':   1500 obs. of  3 variables:
 $ y           : num  0.0125 0.0678 0.1084 0.1757 0.2034 ...
 $ cluster_type: chr  "cluster1" "cluster1" "cluster1" "cluster1" ...
 $ cluster_id  : int  54 10 35 41 96 29 26 48 80 76 ...
> ggplot(pop_long, aes(x = cluster_id, y = y)) + geom_point() +
+     facet_grid(rows = vars(cluster_type))
```

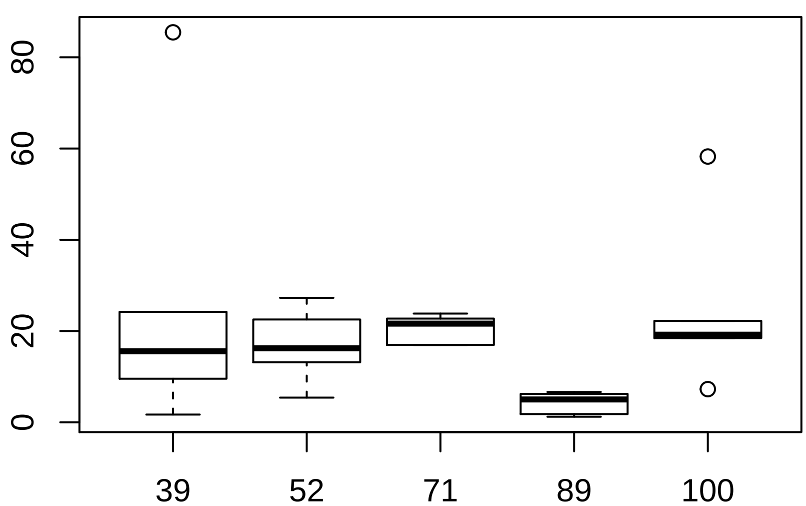


Q1: Consider taking a SRS of $n = 5$ of these clusters and observing all element responses within the cluster. Which choice of cluster variable (`cluster1`, `cluster2` or `cluster3`) will yield a cluster sample that is most like a SRS of 25 elements? Which choice will yield a cluster sample that is least like a SRS of 25 elements?

3. One-stage Cluster Sample

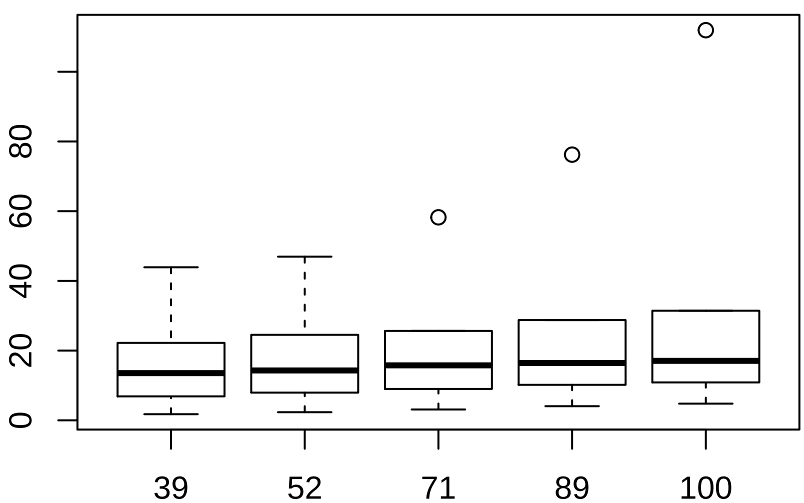
What if we used the `cluster1` variable to define our clusters? Here we sample $n = 5$ cluster ID's and extract the responses

```
> SRS_clusID <- sample(1:100, size = 5, replace = FALSE)
> data_cluster1 <- pop %>% filter(cluster1 %in% SRS_clusID) %>%
+     select(y, cluster1)
> data_cluster1 %>% arrange(cluster1)
           y cluster1
1   1.693726       39
2   9.547130       39
3  15.554906       39
4  24.206683       39
5  85.461437       39
6   5.408245       52
7  13.145772       52
8  16.205982       52
9  22.530824       52
10 27.289074       52
11 16.947582       71
12 16.960375       71
13 21.623214       71
14 22.727779       71
15 23.830603       71
16  1.217529       89
17  1.801813       89
18  5.005002       89
19  6.218676       89
20  6.641853       89
21  7.269267      100
22 18.446701      100
23 19.199765      100
24 22.223270      100
25 58.243358      100
> boxplot(y ~ cluster1, data_cluster1)
```



Similar for using `cluster2` (we can reuse the sample sample of cluster IDs since all three cluster variables just use integers 1-100 to ID clusters):

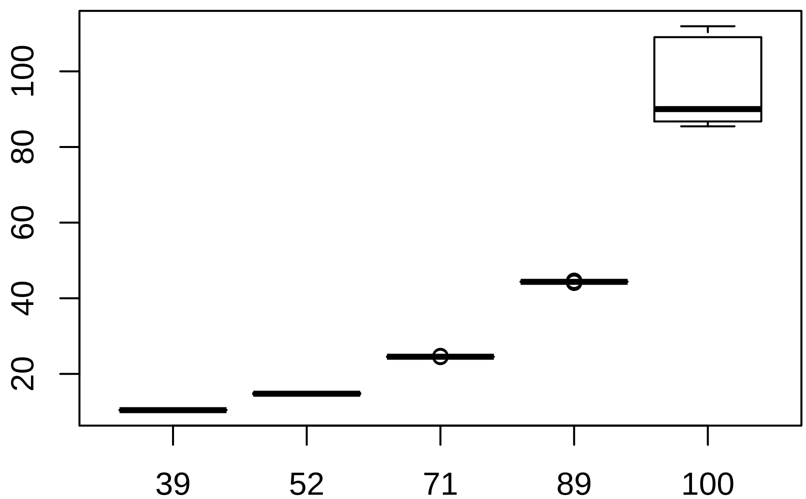
```
> data_cluster2 <- pop %>% filter(cluster2 %in% SRS_clusID) %>%
+     select(y, cluster2)
> data_cluster2 %>% arrange(cluster2)
            y cluster2
1    1.755432       39
2    6.876793       39
3   13.538555       39
4   22.223270       39
5   43.911564       39
6    2.326055       52
7    7.942356       52
8   14.303642       52
9   24.529383       52
10  46.950279       52
11   3.112164       71
12   9.005409       71
13  15.775040       71
14  25.658664       71
15  58.243358       71
16   4.052763       89
17  10.184936       89
18  16.441885       89
19  28.764532       89
20  76.223720       89
21   4.782427      100
22  10.876472      100
23  17.068464      100
24  31.439692      100
25 111.928158      100
> boxplot(y ~ cluster2, data_cluster2)
```



Similar for using `cluster3`:

```
> data_cluster3 <- pop %>% filter(cluster3 %in% SRS_clusID) %>%
+     select(y, cluster3)
> data_cluster3 %>% arrange(cluster3)
           y cluster3
1   10.38140       39
```

```
2   10.38805       39
3   10.40717       39
4   10.48571       39
5   10.55361       39
6   14.60906       52
7   14.61668       52
8   14.75565       52
9   14.96054       52
10  15.03714       52
11  24.48272       71
12  24.52938       71
13  24.53441       71
14  24.54811       71
15  24.71480       71
16  44.09305       89
17  44.34053       89
18  44.34504       89
19  44.43401       89
20  44.61228       89
21  85.46144      100
22  86.75496      100
23  90.01709      100
24 109.04558      100
25 111.92816      100
> boxplot(y ~ cluster3, data_cluster3)
```



Q2: Are these samples of 5 clusters similar reflections on how y does, or does not, depend on cluster ID for the three types of clustering variable?

4. Simulation

Let's repeat part 3. samples many, many times and construct a one-stage cluster estimate of population mean for each. We will also take a SRS of 25 elements and get a SRS estimate of population mean too. For each sample, save the SRS estimate of population mean and the equal-cluster size one-stage estimate of population mean (just the sample mean of all elements).

```
> reps <- 10000  # simulation size
> n <- 5  # cluster sample size
> results <- data.frame(run = 1:reps, est_srs = NA, est_cluster1 = NA,
+     est_cluster2 = NA, est_cluster3 = NA)
>
> for (i in 1:reps) {
+     # SRS
+     SRS_elemID <- sample(1:nrow(pop), size = n * 5, replace = F)  # srs units
+     data_SRS <- pop[SRS_elemID, ]
+     results$est_srs[i] <- mean(data_SRS$y)  # sample mean from SRS
+
+     # cluster sample ID's
+     SRS_clusID <- sample(1:100, size = n, replace = FALSE)
+
+     # cluster sample 1
+     data_cluster1 <- pop %>% filter(cluster1 %in% SRS_clusID)
+     results$est_cluster1[i] <- sum(data_cluster1$y)/(5 * n)  # unbiased/ratio
+
+     # cluster sample 2
+     data_cluster2 <- pop %>% filter(cluster2 %in% SRS_clusID)
+     results$est_cluster2[i] <- sum(data_cluster2$y)/(5 * n)  # unbiased/ratio
+
+     # cluster sample 3
+     data_cluster3 <- pop %>% filter(cluster3 %in% SRS_clusID)
+     results$est_cluster3[i] <- sum(data_cluster3$y)/(5 * n)  # unbiased/ratio
+
+ }
```

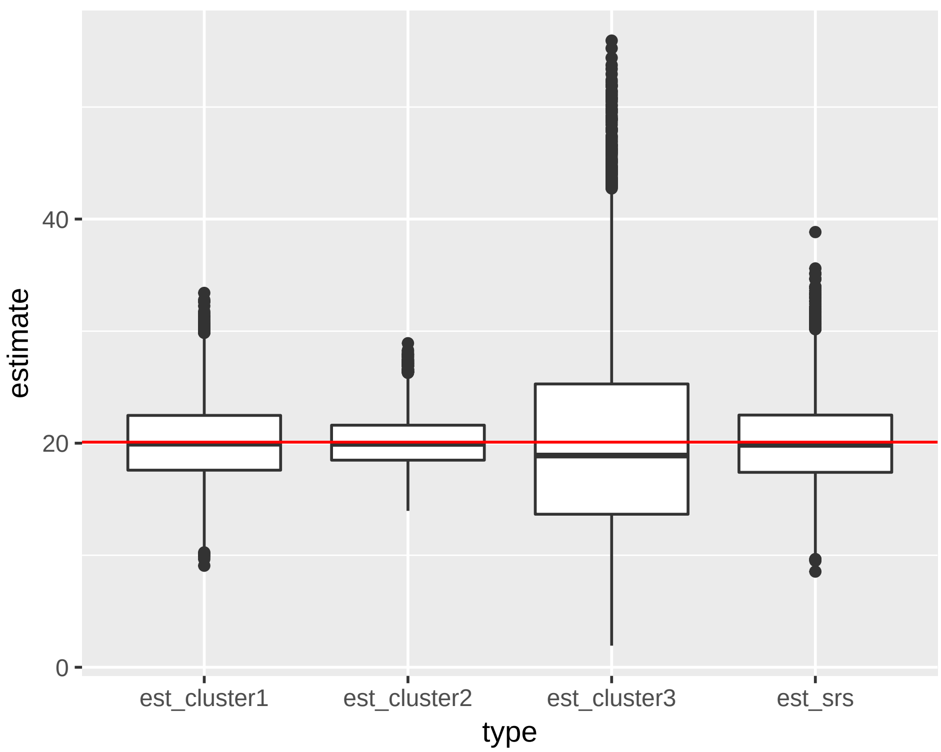
5. Compare Sampling Distributions

The population mean is just over 20.

```
> pop_mean <- mean(pop$y)
> pop_mean
[1] 20.09308
```

How do our estimators compare in terms of bias and variability? We can make a boxplot of simulated sampling distributions of our four types of estimators:

```
> str(results)
'data.frame':   10000 obs. of  5 variables:
 $ run         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ est_srs     : num  26.9 16.8 19.4 22 24 ...
 $ est_cluster1: num  17.3 19.1 24.9 17.8 16.5 ...
 $ est_cluster2: num  22.4 17.5 19.2 18.3 18.1 ...
 $ est_cluster3: num  26.1 11.5 16.5 12.6 12.5 ...
> results_long <- results %>% gather(key = type, value = estimate,
+     starts_with("est"))
> str(results_long)
'data.frame':   40000 obs. of  3 variables:
 $ run     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ type    : chr  "est_srs" "est_srs" "est_srs" "est_srs" ...
 $ estimate: num  26.9 16.8 19.4 22 24 ...
> ggplot(results_long, aes(x = type, y = estimate)) + geom_boxplot() +
+     geom_hline(yintercept = pop_mean, color = "red")
```



And we can get simulated bias and SE:

```
> results_long %>% group_by(type) %>% summarize(expected_value = mean(estimate),
+     bias = expected_value - pop_mean, percent_bias = 100 * bias/pop_mean,
```

```
+     SE = sd(estimate))
# A tibble: 4 x 5
  type         expected_value     bias percent_bias    SE
  <chr>                 <dbl>    <dbl>        <dbl> <dbl>
1 est_cluster1           20.1 -0.0240      -0.119    3.55
2 est_cluster2           20.1  0.0132       0.0657   2.27
3 est_cluster3           20.1  0.00184      0.00916  8.52
4 est_srs                20.1 -0.0413      -0.206    3.78
> pop_mean
[1] 20.09308
```

Q3 (goal 1) For a clustering variable `cluster1`, compare precision of a one stage cluster sample of $n = 5$ clusters (with $nM = 25$ elements) to a SRS of $n = 25$ elements.

Q4 (goal 2) How does **Q3** depend on the clustering variable? COmpare the SRS to the choice of `cluster2` and `cluster3`. When will a cluster sample "beat" a SRS? WHen does a SRS "beat" a cluster sample? When are they similar? Think about how to write down a general rule of thumb for when cluster sampling is better than a SRS, in terms of precision.