Report 6
Systematic Sampling

• **Name of the project:** Analysis of Average Winter Temperature in College Park, Maryland

Variables of Interest

• The true parameter: the true average temperature (in ℉) in College Park, MD during winter months, **μ** = 39.177

• The population size: *M* = 720

• The total sample size from Report 2: *n* = 84

1. *Perform a systematic sample with (approximately) the same total sample size as you used in previous Reports. For example, your population size M=300 and total sample size is 60. The systematic sample can be taken in many ways. One of them is to choose the first N=10 units. Out of these units take 2 by SRS. Suppose units 3 and 7 are selected. Then from next ten units we select two by this pattern. And so on. The other way is to choose the first N=15 units and select from them 3 units by SRS. Then repeat this pattern on next 15 and so on.*

With a population size of M=720, our target sample size is n=84. We can be conservative and take a slightly larger sample of size n=90. This allows us to simply take a **2-in-16 sample**:

- Take the first N=16 units.
- Out of these units, randomly choose two by SRS.
- Choose the units in those same 2 positions, for each consecutive 16 units.

Why not a "1-in-8" sample? Because then we would be using a single systematic sampling design, which we cannot reliably calculate variance for.

Code:

```
winterData = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")

temp = winterData$temp

M = 720 #total number of secondary units

N = 16 #num of systematic samples to choose from
```

```
n = 2 #num of systematic samples we will choose
#the sample follows an "n-in-N" design
#how many primary units do we cycle through?
primaries = M/N

group1 = c()
group2 = c()
#the positions we will draw from
positions = sample(1:N, size=n)
for(i in 1:primaries) {
  for(j in 1:length(positions)) {
    if (j == 1){
      index = (i-1)*N + positions[j]
      group1 = append(group1, temp[index])

    } else if (j == 2) {
      index = (i-1)*N + positions[j]
      group2 = append(group2, temp[index])
    }
  }
}
group1
group2
sample = c(group1, group2)
```

Output:

```
> group1
 [1] 46.0 30.5 35.6 33.1 19.8 41.0 48.3 28.4 23.5 36.3 41.5 32.9 56.6 61.1
38.5 40.4 48.3 39.7 16.9 29.5 25.4 42.7 36.4 42.1 46.7 48.9
[27] 30.8 43.5 37.0 48.3 45.4 34.0 27.7 42.2 30.9 39.4 37.4 44.5 30.3 36.2
57.2 33.5 26.2 48.9 43.1
> group2
 [1] 40.9 37.1 25.5 38.3 35.9 30.5 51.8 59.4 35.1 31.4 20.5 52.7 34.3 42.6
```

```
50.0 34.5 61.0 43.1 53.1 27.9 41.9 34.8 49.3 43.0 47.7 28.9

[27] 22.1 55.2 41.0 26.4 52.5 27.8 44.5 30.4 35.5 35.1 39.1 34.0 33.4 42.6
45.7 52.9 22.5 31.0 57.9
```

This particular sample used every element in the 15th and 5th position.

2. *Estimate your parameter of interest by an unbiased estimator. Estimate its variance and give a confidence interval of α level chosen in Report 2.*

Unbiased Estimator: $\hat{\mu} = \hat{\tau}/M = \frac{N}{n*M}\sum_{i=1}^{n} y_i = \frac{N\bar{y}}{M}$ = 38.97111

where N= # of primary units in population (16), n = # of primary units in sample (2), M= # of secondary units (720).

Code:

```
muhat = N*sum(sample) / (n*M)

muhat

[1] 38.97111
```

Estimated variance of this estimator:

$\hat{var}(\hat{\mu}) = \frac{\hat{var}(\hat{\tau})}{M^2} = N(N-n)\frac{s_u^2}{nM^2}$ = 0.3173377

where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, n is # of units randomly selected

Code:

```
sums = c(sum(group1), sum(group2))

su2 = var(sums)


varhattauhat = N*(N-n)*(su2/n)

varhatmuhat = varhattauhat/(M^2)
```

```
varhatmuhat

[1] 0.3173377
```

Confidence Interval:

$CI95 = \hat{\mu} \pm t_{\alpha/2, n-1}\sqrt{\widehat{var}(\hat{\mu})}$ = (31.81336, 46.12886)

Width: 14.3155

Code:

```
# df = 2-1 = 1

CI95 = c(muhat - qt(.975, n-1)*sqrt(varhatmuhat), muhat + qt(.975,
n-1)*sqrt(varhatmuhat))

CI95

[1] 31.81336 46.12886
```

3. *Take another systematic sample.*

We'll perform a **3-in-24 sample** in the same way we did before. The code is similar to before, but now N = 24 and n = 3 (so now the sample gets divided into three groups instead of two).

```
winterData = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")

temp = winterData$temp

M = 720 #total number of secondary units

N = 24 #num of systematic samples to choose from

n = 3 #num of systematic samples we will choose

#the sample follows an "n-in-N" design

#how many primary units do we cycle through?

primaries = M/N


group1 = c()

group2 = c()
```

```
group3 = c()

#the positions we will draw from

positions = sample(1:N, size=n)

for(i in 1:primaries) {

  for(j in 1:length(positions)) {

    if (j == 1){

      index = (i-1)*N + positions[j]

      group1 = append(group1, temp[index])


    } else if (j == 2) {

      index = (i-1)*N + positions[j]

      group2 = append(group2, temp[index])

    } else if (j == 3) {

      index = (i-1)*N + positions[j]

      group3 = append(group3, temp[index])

    }

  }

}

group1

group2

group3

sample = c(group1, group2, group3)
```

Output:

```
> group1

 [1] 43.6 30.2 36.5 32.6 56.0 42.6 39.4 41.4 43.3 47.8 37.4 34.6 22.8 49.2
62.2 44.8 42.8 26.6 40.9 34.1 48.9 37.0 48.5 43.0 38.4 32.3

[27] 45.1 31.4 26.2 30.8

> group2

 [1] 37.9 29.1 39.2 23.4 66.4 34.4 30.9 42.1 37.1 43.8 35.0 39.4 14.7 43.9
52.9 39.6 31.0 43.2 41.2 41.0 32.6 45.3 39.6 37.7 32.0 37.2

[27] 43.5 34.7 24.4 56.2

> group3
```

```
 [1] 37.1 28.1 35.9 45.6 59.4 23.5 20.5 32.8 42.6 50.5 61.0 27.7 27.9 44.9
49.3 60.0 28.9 48.9 41.0 46.0 27.8 47.3 35.5 44.1 34.0 33.2

[27] 45.7 26.1 31.0 39.9
```

This particular sample used every 16th, 18th, and 21st element.

4. *Estimate your parameter of interest by an unbiased estimator. Estimate its variance and give a confidence interval of α level chosen in Report 2.*

Unbiased Estimator: $\hat{\mu} = \hat{\tau}/M = \frac{N}{n*M}\sum_{i=1}^{n} y_i = \frac{N\bar{y}}{M}$ = 39.06667

where N= # of primary units in population (24), n= # of primary units in sample (3), M= # of secondary units (720).

Code:

```
muhat = N*sum(sample) / (n*M)

muhat

[1] 39.06667
```

Estimated variance of this estimator:

$\hat{var}(\hat{\mu}) = \frac{\hat{var}(\hat{\tau})}{M^2} = N(N-n)\frac{s_u^2}{nM^2}$ = 0.2723843

where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, n is # of units randomly selected

Code:

```
sums = c(sum(group1), sum(group2), sum(group3))

su2 = var(sums)


varhattauhat = N*(N-n)*(su2/n)

varhatmuhat = varhattauhat/(M^2)

varhatmuhat

[1] 0.2723843
```

Confidence Interval:

$CI95 \ = \hat{\mu} \ \pm \ t_{\alpha/2,\, n-1}\sqrt{\widehat{var}(\hat{\mu})}$ = (36.82110, 41.31224)

Width: 4.49114

Code:

```
# df = 3-1 = 2
```

```
CI95 = c(muhat - qt(.975, n-1)*sqrt(varhatmuhat), muhat + qt(.975,
n-1)*sqrt(varhatmuhat))

CI95

[1] 36.82110 41.31224
```

5. Order your population with respect to y (variable of interest). Repeat steps 1-4 for the ordered data.

Sort the data:

```
temp = sort(winterData$temp)
```

Here, we re-used the code in steps 1-4, with the only difference being that we used the sorted temp variable when performing the systematic sample.

<u>First systematic sample: **2-in-16**</u>

Positions chosen: Every 7th and every 14th

Sample:

```
> group1

 [1] 14.8 20.3 23.0 25.2 27.2 28.6 29.6 30.5 31.2 32.0 32.7 33.4 34.0 34.6
35.0 35.5 36.1 36.4 37.1 37.4 38.1 38.4 39.0 39.6 40.2 40.6

[27] 41.0 41.4 42.1 42.6 43.1 43.7 44.3 44.9 45.5 46.2 46.9 47.9 48.9 49.7
51.0 52.9 55.3 58.0 61.6

> group2
```

```
 [1] 18.3 21.9 23.8 26.1 28.0 29.1 30.2 30.7 31.5 32.3 33.1 33.5 34.2 34.7
35.2 35.7 36.3 36.8 37.3 37.7 38.3 38.6 39.3 39.8 40.3 40.9

[27] 41.2 41.7 42.3 42.8 43.4 44.0 44.7 45.2 45.8 46.5 47.3 48.3 49.3 50.1
51.8 53.1 56.0 59.4 64.8
```

Calculations:

Unbiased estimator: $\widehat{\mu} = 39.32$

Estimated variance: $\widehat{var}(\widehat{\mu}) = 0.06118951$

CI95: $CI95 = (36.17693,\ 42.46307)$

Width: 6.28614

Second systematic sample: **3-in-24**

Positions chosen: Every 4th, 18th and 23rd

Sample:

```
> group1

 [1] 13.8 21.8 24.8 27.8 29.5 30.7 31.9 33.0 34.0 34.7 35.4 36.2 37.0 37.6
38.4 39.2 40.1 40.8 41.3 42.2 43.0 43.9 44.8 45.7 46.8 48.2

[27] 49.5 51.6 54.9 58.2

> group2

 [1] 19.7 23.5 26.6 28.8 30.4 31.3 32.4 33.4 34.4 35.1 35.9 36.5 37.3 38.2
38.7 39.7 40.4 41.1 41.9 42.7 43.6 44.5 45.3 46.3 47.6 49.2

[27] 50.3 53.0 56.9 62.1

> group3

 [1] 20.3 23.9 27.2 29.1 30.5 31.8 32.7 33.6 34.6 35.3 36.1 36.8 37.4 38.3
39.0 39.9 40.6 41.3 42.1 42.8 43.7 44.7 45.5 46.6 47.9 49.3

[27] 51.0 53.3 58.0 66.4
```

Calculations:

Unbiased estimator: $\widehat{\mu} = 39.37$

Estimated variance: $\widehat{var}(\widehat{\mu}) = 0.1458333$

CI95: $CI95 = (37.7269, 41.0131)$

Width: 3.2862

6. *Choose the best estimator of your parameter based on the width of CI.*

Widths of the estimators' confidence intervals are as follows:

2-in-16, unsorted: w = 14.3155

3-in-24, unsorted: w = 4.49114

2-in-16, sorted: w = 6.28614

3-in-24, sorted: w = **3.2862**

Thus, the best estimator is obtained from using the 3-in–24 systematic sampling with the SORTED dataset since the width of the CI is the narrowest among the four CIs.

Looking at the results above, we can see that performing the systematic sampling with the sorted dataset yielded more narrow CIs than their respective systematic sampling on the unsorted dataset (ie 2-in-16 sorted narrower CI than 2-in-16 unsorted and same for 3-in-24). Comparing the 2-in-16 versus the 3-in-24 systematic sampling, the 3-in-24 systematic sampling yielded more narrow CIs compared to the 2-in-16 systematic samplings.

7. *Show all formulas used at each step as well as the code.*

Done above.