Report 7
Two Stage Sampling.

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

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

The population size *M* = 720

The total sample size *n* = 84

1. Divide your population into *N* (3 to 5) primary units. You can use any variable from the data set. It does not have to be one of the two which were used as auxiliary variables or for stratification.

   We will divide our population in **N = 4 primary units** organized by year as follows:

   To calculate mi for each of the years, we will perform a simple random sample with n = 3 so that mi = 84/3 = 28.

Years	Mi
2014/2015/2016	211
2017/2018	180
2019/2020	180
2021/2022	149
[total]	720

Code:

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

temp = df[1:720, 5]

dew = df[1:720, 9]

date = df[1:720, 2]


group1temp = c()

group2temp = c()

group3temp = c()
```

```
group4temp = c()

#build the primary units based on the month year of datetime.
for(i in 1:length(date)){
  yr = substr(date[i], nchar(date[i])-1, nchar(date[i]))
  if(yr == "14" || yr == "15" || yr == "16") {

    group1temp = append(group1temp, temp[i])
  } else if (yr == "17" || yr == "18") {
    group2temp = append(group2temp, temp[i])
  } else if (yr == "19" || yr == "20") {
    group3temp = append(group3temp, temp[i])
  } else if (yr == "21" || yr == "22") {
    group4temp = append(group4temp, temp[i])
  }

}
M1 = length(group1temp)
M2 = length(group2temp)
M3 = length(group3temp)
M4 = length(group4temp)

> M1
[1] 211
> M2
[1] 180
> M3
```

```
[1] 180

> M4

[1] 149
```

2. Perform two stage design with SRS at each stage with total size $\sum_{i=1}^{n} m_i$ chosen in Report 2.

We have N=4 primary units and will use n=3. From those 3 selected primary units, we will choose mi=28 from each.

Code:

```
#select which three groups will be used

groups = sample(1:4, 3, replace=F)

groups #groups selected

[1] 3 4 2


Mi = c()

yibar = c()

sisq = c()


#select 28 values from each group in groups

if (1 %in% groups) {

  samp1 = sample(1:M1, 28, replace=F) #indices of our sample

  samp1temp = group1temp[samp1]

  Mi = append(Mi, 211)

  yibar = append(yibar, mean(samp1temp))

  sisq = append(sisq, var(samp1temp))

}

if (2 %in% groups) {

  samp2 = sample(1:M2, 28, replace=F)#indices of our sample
```

```
  samp2temp = group2temp[samp2]
  Mi = append(Mi, 180)
  yibar = append(yibar, mean(samp2temp))
  sisq = append(sisq, var(samp2temp))
}
if (3 %in% groups) {
  samp3 = sample(1:M3, 28, replace=F)#indices of our sample
  samp3temp = group3temp[samp3]
  Mi = append(Mi, 180)
  yibar = append(yibar, mean(samp3temp))
  sisq = append(sisq, var(samp3temp))
}
if (4 %in% groups) {
  samp4 = sample(1:M4, 28, replace=F)#indices of our sample
  samp4temp = group4temp[samp4]
  Mi = append(Mi, 149)
  yibar = append(yibar, mean(samp4temp))
  sisq = append(sisq, var(samp4temp))
}

Mi
[1] 180 180 149
yibar
[1] 41.72500 37.70000 39.92857
> samp1temp
NULL
> samp2temp
```

```
 [1] 61.6 44.1 44.7 46.8 41.4 43.8 33.6 31.9 41.3 55.8 46.6
27.2 35.0 43.9 36.0 43.0 22.5 40.6 41.1 44.9 39.6 62.1 49.2
45.4 33.3 24.9

[27] 42.3 45.7

> samp3temp

 [1] 41.4 38.6 46.9 33.5 42.4 30.3 29.7 41.2 34.0 44.5 23.0
42.7 40.6 41.8 30.8 51.9 46.0 34.1 34.6 30.1 44.0 29.3 40.9
41.3 30.6 28.0

[27] 40.4 43.0

> samp4temp

 [1] 36.9 48.9 49.3 46.9 33.2 49.8 34.7 44.3 26.2 40.2 46.8
38.5 30.3 39.9 42.6 47.1 37.9 36.8 35.3 30.7 39.1 53.0 40.1
34.0 39.7 34.8

[27] 43.9 37.1
```

3. Estimate your parameter of interest by unbiased estimator. Estimate its variance and standard deviation.

- $\hat{\mu} = \hat{\tau}/M = \frac{1}{M}\left(\frac{N}{n}\sum_{i=1}^{n}\hat{y}_i\right) = \frac{1}{M}\left(\frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_i\right) =$ **37.49233**
- $\widehat{var}(\hat{\mu}) = \widehat{var}(\hat{\tau})/M^2 = \frac{1}{M^2}\left(N(N-n)\frac{s_u^{\,2}}{n} + \frac{N}{n}\sum_{i=1}^{n}M_i(M_i - m_i)\frac{s_i^{\,2}}{m_i}\right) =$ **1.98372**

Where… $s_u^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \hat{\mu}_1\right)^2$; $s_i^{\,2} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$

- Estimated Standard Deviation = **1.408446**

Code:

```
#unbiased estimator

mi = c(28, 28, 28)

M = 720

N = 4

n = 3
```

```
yihat = Mi*yibar

muhat = N*mean(yihat)/M

muhat

[1] 37.49233


varhatmuhat = (N*(N-n)*var(yihat)/n +N*sum(Mi*(Mi-mi)*sisq/(n*mi)))/M^2

varhatmuhat

[1] 1.98372

sqrt(varhatmuhat)

[1] 1.408446
```

4. Estimate your parameter of interest by ratio estimator. Estimate its variance and standard deviation.

- $\hat{\mu}_r = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i} = \hat{r}$ = **39.77575**
- $\widehat{var}(\hat{\mu}_r) = \widehat{var}(\hat{\tau}_r) / M^2 = \frac{1}{M^2}\left(\frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} \left(\hat{y}_i - M_i \hat{r}\right)^2 + \frac{N}{n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{s_i^2}{m_i}\right)$ = **0.7524291**

  Where… s_i^2 is the same as calculated in the prior step.

- Estimated Standard Deviation = **0.8674267**

Code:

```
rhat = sum(yihat)/sum(Mi)

> rhat

[1] 39.77575


varhatrhat =
(N*(N-n)*var(yihat-Mi*rhat)/n+N*sum(Mi*(Mi-mi)*sisq/(n*mi)))/M^2

> varhatrhat

[1] 0.7524291
```

```
> sqrt(varhatrhat)
[1] 0.8674267
```

5. Perform two stage design in which primary units are selected with replacement, with probabilities proportional to size and a sample of secondary units is selected independently using SRS with the total size $P^{n}_{i=1} m_i$ chosen in Report 2.

Code:

```
allMis = c(M1, M2, M3, M4)
pis = allMis/sum(allMis)
> pis
[1] 0.2930556 0.2500000 0.2500000 0.2069444

groups = sample(1:4, 3, replace=T, prob=pis)
> groups #groups selected
[1] 2 4 3

fullyibar = c()
#creates sample of 28 values for each group
samp1 = sample(1:M1, 28, replace=F) #indices of our sample
samp1temp = group1temp[samp1]
fullyibar = append(fullyibar, mean(samp1temp))

samp2 = sample(1:M2, 28, replace=F)#indices of our sample
samp2temp = group2temp[samp2]
fullyibar = append(fullyibar, mean(samp2temp))

samp3 = sample(1:M3, 28, replace=F) #indices of our sample
samp3temp = group3temp[samp3]
fullyibar = append(fullyibar, mean(samp3temp))
```

```
samp4 = sample(1:M4, 28, replace=F) #indices of our sample

samp4temp = group4temp[samp4]

fullyibar = append(fullyibar, mean(samp4temp))

> fullyibar

[1] 36.88929 37.30357 41.44643 38.25714


yibar = c()

for(i in 1:length(groups)) {

  if (groups[i] == 1) {

    yibar = append(yibar, fullyibar[1])

  }

  if (groups[i] == 2) {

    yibar = append(yibar, fullyibar[2])

  }

  if (groups[i] == 3) {

    yibar = append(yibar, fullyibar[3])

  }

  if (groups[i] == 4) {

    yibar = append(yibar, fullyibar[4])

  }

}



> group2temp[samp2]

 [1] 22.6 24.9 44.9 27.2 36.3 35.5 52.2 26.1 34.0 45.4 31.0 36.4 39.6
34.4 42.8 50.0 20.1 51.9 49.2 16.9 39.8 63.9 36.6 50.0 36.1 39.4

[27] 22.7 34.6

> group4temp[samp4]
```

```
 [1] 29.7 47.6 34.3 54.3 37.1 43.5 27.0 39.4 36.9 44.1 44.3 33.2 28.6
31.0 40.3 26.2 38.4 47.1 38.3 34.0 22.5 39.1 29.2 45.7 46.5 40.9

[27] 34.0 58.0

> group3temp[samp3]

 [1] 41.3 34.6 34.0 47.6 41.0 34.0 41.3 41.2 41.4 40.7 42.2 46.0 29.6
53.3 46.0 46.3 39.2 37.0 42.2 40.6 44.5 38.3 49.2 46.2 49.5 28.0

[27] 45.2 40.1


yibar

[1] 37.30357 38.25714 41.44643
```

6. Estimate your parameter of interest by Hansen-Horvitz estimator. Estimate its variance and standard deviation.

- $\hat{\mu}_p = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} =$ **39.00238**
- $\widehat{var}(\hat{\mu}_p) = \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \left(\bar{y}_i - \hat{\mu}_p \right)^2 \right) =$ **1.569117**
- Estimated Standard deviation = **1.252644**

Code:

```
> yibar

[1] 37.30357 38.25714 41.44643


muhat = mean(yibar)

> muhat

[1] 39.00238


varhatmuhat = var(yibar)/n

> varhatmuhat

[1] 1.569117
```

```
> sqrt(varhatmuhat)
[1] 1.252644
```

7. Divide your population into *N* primary units in a different way from part 1.You can use any variable from the data set. It does not have to be one of the two which were used as auxiliary variables or for stratification.

We will now divide our population in **N = 3 primary units** organized by month as follows:

To calculate mi for each of the months, we will perform a simple random sample with n = 2 so that mi = 84/2 = 42.

Month	Mi
December	248
January	248
February	224
[total]	720

```
df = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")
temp = df[1:720, 5]
dew = df[1:720, 9]
date = df[1:720, 2]

group1temp = c() #Dec
group2temp = c() #Jan
group3temp = c() #Feb

#build the primary units based on the month value of datetime.
for(i in 1:length(date)){
```

```
  month = substr(date[i], 1, 2)
  if(month == "12") {
    group1temp = append(group1temp, temp[i])
  } else if (month == "1/") {
    group2temp = append(group2temp, temp[i])
  } else if (month == "2/") {
    group3temp = append(group3temp, temp[i])
  }
}

mi = 42
M1 = length(group1temp)
M2 = length(group2temp)
M3 = length(group3temp)

> M1
[1] 248
> M2
[1] 248
> M3
[1] 224
```

8. Repeat steps 2-6.

(1) Two stage design with SRS at each stage…

In report 2, we settled for a sample size of 84. Since there are 3 primary units, it only makes sense to sample 2 of them. If we draw 42 of each (mi = 42), we get our desired total sample size of 84.

Code:

```
samp1temp = c() #Dec

samp2temp = c() #Jan

samp3temp = c() #Feb


#select which two groups will be used

groups = sample(1:3, 2, replace=F)

> groups #groups selected for sampling

[1] 1 3


Mi = c()

yibar = c()

sisq = c()


#select 42 values from each group in groups

if (1 %in% groups) { #Dec

  samp1 = sample(1:M1, 42, replace=F) #indices of our sample

  samp1temp = group1temp[samp1]

  Mi = append(Mi, 248)

  yibar = append(yibar, mean(samp1temp))

  sisq = append(sisq, var(samp1temp))

}

if (2 %in% groups) { #Jan
```

```
  samp2 = sample(1:M2, 42, replace=F) #indices of our sample
  samp2temp = group2temp[samp2]
  Mi = append(Mi, 248)
  yibar = append(yibar, mean(samp2temp))
  sisq = append(sisq, var(samp2temp))
}
if (3 %in% groups) { #Feb
  samp3 = sample(1:M3, 42, replace=F) #indices of our sample
  samp3temp = group3temp[samp3]
  Mi = append(Mi, 224)
  yibar = append(yibar, mean(samp3temp))
  sisq = append(sisq, var(samp3temp))
}
> yibar
[1] 40.40476 38.14048


> samp1temp
 [1] 41.8 34.8 41.1 39.6 29.2 30.6 43.0 32.9 35.9 41.0 46.0 58.2 18.3 33.0 43.5
48.5 44.7 53.0 38.4 52.9 39.8 30.9 46.2 42.2 37.4
[26] 31.5 35.5 41.2 38.6 36.5 56.6 37.4 60.5 27.2 26.0 42.0 39.9 64.8 40.3 27.7
47.9 40.5
> samp2temp #42 chosen from Jan
NULL
> samp3temp #42 chosen from Feb
 [1] 38.4 39.4 23.9 55.2 35.6 28.8 48.4 34.4 43.2 36.3 43.1 41.0 31.2 37.3 47.6
28.4 61.3 39.5 31.9 51.2 41.3 36.1 37.2 39.2 23.4
[26] 21.3 55.6 51.3 29.1 46.6 33.2 47.7 19.9 57.0 20.5 30.4 13.8 48.1 49.4 48.9
19.7 36.1
```

Note also that February is slightly overrepresented when it is selected, since there are only 224 total entries for February as opposed to 248 for each December and January.

(2) Estimate mean temp with unbiased estimator…

- $\hat{\mu} = \hat{\tau}/M = \frac{1}{M}\left(\frac{N}{n}\sum_{i=1}^{n}\hat{y}_i\right) = \frac{1}{M}\left(\frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_i\right) =$ **38.67468**
- $\widehat{var}(\hat{\mu}) = \widehat{var}(\hat{\tau})/M^2 = \frac{1}{M^2}\left(N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n}M_i(M_i - m_i)\frac{s_i^2}{m_i}\right) =$ **3.846573**

Where… $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \hat{\mu}_1\right)^2$; $s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$

- Estimated Standard Deviation = **1.961268**

Code:

```
mi = c(42, 42)

M = 720

N = 3

n = 2

yihat = Mi*yibar

> yihat

[1] 10020.381  8543.467


muhat = N*mean(yihat)/M

> muhat

[1] 38.67468


varhatmuhat = (N*(N-n)*var(yihat)/n +N*sum(Mi*(Mi-mi)*sisq/(n*mi)))/M^2

> varhatmuhat

[1] 3.846573


> sqrt(varhatmuhat)

[1] 1.961268
```

(3) Estimate mean temp with ratio estimator…

- $\hat{\mu}_r = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i} = \hat{r}$ = **39.33019**
- $\widehat{var}(\hat{\mu}_r) = \widehat{var}(\hat{\tau}_r) / M^2 = \frac{1}{M^2}\left(\frac{N(N-n)}{n(n-1)}\sum_{i=1}^{n}\left(\hat{y}_i - M_i\hat{r}\right)^2 + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}\right)$ = **1.101784**

  Where… s_i^2 is the same as calculated in the prior step.

- Estimated Standard Deviation = **1.049659**

Code:

```
rhat = sum(yihat)/sum(Mi)

> rhat

[1] 39.33019


> varhatrhat =(N*(N-n)*var(yihat-Mi*rhat)/n
+N*sum(Mi*(Mi-mi)*sisq/(n*mi)))/M^2

> varhatrhat

[1] 1.101784


> sqrt(varhatrhat)

[1] 1.049659
```

(4) Sampling with PPS for the primary units…

This won't be much different from the results we already observed. The only key difference is that February has slightly less entries, so it's slightly less likely to be chosen in the sample of primary units.

```
allMis = c(M1, M2, M3)

pis = allMis/sum(allMis)

> pis

[1] 0.3444444 0.3444444 0.3111111


groups = sample(1:3, 2, replace=T, prob=pis)

> groups

[1] 3 2


fullyibar = c()


#creates sample of 42 values for each group

samp1 = sample(1:M1, 42, replace=F) #indices of our sample

samp1temp = group1temp[samp1]

fullyibar = append(fullyibar, mean(samp1temp))


samp2 = sample(1:M2, 42, replace=F) #indices of our sample

samp2temp = group2temp[samp2]

fullyibar = append(fullyibar, mean(samp2temp))


samp3 = sample(1:M3, 42, replace=F) #indices of our sample

samp3temp = group3temp[samp3]

fullyibar = append(fullyibar, mean(samp3temp))
```

```
> fullyibar
[1] 43.38810 36.65238 39.05238


yibar = c()


for(i in 1:length(groups)) {
  if (groups[i] == 1) {
    yibar = append(yibar, fullyibar[1])
  }
  if (groups[i] == 2) {
    yibar = append(yibar, fullyibar[2])
  }
  if (groups[i] == 3) {
    yibar = append(yibar, fullyibar[3])
  }
}


> group3temp[samp3]
 [1] 45.7 47.6 33.3 19.8 29.9 37.3 37.6 26.6 58.2 47.2 30.8 43.1 49.6
21.3 32.4
[16] 32.6 47.1 61.6 21.4 49.2 37.7 27.0 29.1 38.4 38.7 33.1 31.9 30.8
48.8 29.5
[31] 35.3 29.5 51.3 39.9 32.6 56.2 43.8 39.5 45.7 49.2 55.2 44.7
> group2temp[samp2]
 [1] 30.4 24.0 23.5 40.0 40.6 42.8 16.9 38.3 46.5 32.0 52.1 25.5 36.4
34.6 46.6
[16] 42.6 34.5 44.6 35.6 32.0 15.2 37.2 30.5 42.8 40.2 55.9 37.4 33.9
40.3 50.5
[31] 27.8 28.6 43.6 25.1 41.0 40.6 28.9 28.1 38.5 52.9 48.9 32.0
```

```
> yibar
[1] 39.05238 36.65238
```

(5) Estimate mean temp with Hansen-Horvitz estimator…

- $\hat{\mu}_p = \frac{\sum_{i=1}^{n} \bar{y}_i}{n}$ = **37.85238**
- $\hat{var}(\hat{\mu}_p) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\mu}_p\right)^2\right)$ = **1.44**
- Estimated Standard Deviation = **1.2**

Code:

```
muhat = mean(yibar)
> muhat
[1] 37.85238

varhatmuhat = var(yibar)/n
> varhatmuhat
[1] 1.44

std = sqrt(varhatmuhat)
> std
[1] 1.2
```

9. Choose the best estimator of your parameter based on the smallest standard deviation (variance).

   The order of variances from each of the estimators from least to greatest is as follows:

Let's look at our results:

Estimator	Variance	Standard Deviation
Unbiased Estimator Using Years	1.98372	1.408446
Ratio Estimator Using Years	0.7524291	0.8674267
Hansen-Horvitz Estimator Using Years	1.569117	1.252644
Unbiased Estimator Using Months	3.846573	1.961268
Ratio Estimator Using Months	1.101784	1.049659
Hansen-Horvitz Estimator Using Months	1.44	1.2

Thus, the best estimator to choose with the smallest variance would be the Ratio Estimator from dividing the population into primary units based on years.

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

    Done at each step.