Report 5
Stratified 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

The population size *N* = 720

The total sample size *n* = 84

1. Divide your population into strata. You can use any variable from the data set. It does not have to be one of two which were used for previous two reports. Follow the stratification principle: stratify the population in such a way that within each stratum the units are as similar as possible.
   To check this, calculate:

$$d = (N-1)\sigma^2 - \sum_{h=1}^{L}(N_h - 1)\sigma_h^2 \text{ where}$$

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2,\ h = 1, \ldots, L$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

Choose stratification (variable for stratification) of the population with largest *d*.

Solution:

There are three different ways to make stratas as follows:

A) We can take dew as the variable for stratification and use the winter months to create 3 strata that divide the population into data from December, January, and February (refer to code A) . The resulting value of d is 4719.88.
B) Another way to create strata for the dew point is stratifying every 10˚F (refer to code B). The value of d for this stratification is 51024.52
C) We also stratified dew point every 20˚F, and got a d of 41389.77 (refer to code C).

In conclusion, we will use the method from part B (stratifying the dew point every 10˚F) since stratification B resulted in the greatest value for d.

Note: I attempted to also try using humidity as an auxiliary variable. However, when using humidity as the auxiliary variable for the latter parts, the confidence interval did not contain the temperature's true mean. This makes sense because the humidity variable is not well correlated with temperature, whereas dew is highly correlated with temperature.

A – Code to stratify dew point by winter months (Dec., Jan., and Feb.):

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

#create strata based on month: December, January, February
tempDec = c(temp[1:31], temp[91:121], temp[181:211], temp[271:301],
temp[361:391], temp[451:481], temp[541:571], temp[631:661])

tempJan = c(temp[32:62], temp[122:152], temp[212:242], temp[302:332],
temp[392:422], temp[482:512], temp[572:602], temp[662:692])

tempFeb = c(temp[63:90], temp[153:180], temp[243:270], temp[333:360],
temp[423:450], temp[513:540], temp[603:630], temp[693:720])

NDec = length(tempDec)
NJan = length(tempJan)
NFeb = length(tempFeb)

> d = (N-1)*var(temp) - sum((c(NDec, NJan, NFeb) - 1)*c(var(tempDec),
var(tempJan), var(tempFeb)))
> d
[1] 4719.88
```

B – Code to stratify dew point every 10˚F:

```
df = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")
temp <- df[1:720, 5]
dew <- df[1:720, 9]
humidity <- df[1:720, 10]

N = 720
n = 84

# creating strata from dew population (every 10)
temp0 = subset(temp, dew < 0)
temp1 = subset(temp, dew >= 0 & dew < 10)
temp2 = subset(temp, dew >= 10 & dew < 20)
temp3 = subset(temp, dew >= 20 & dew < 30)
temp4 = subset(temp, dew >= 30 & dew < 40)
temp5 = subset(temp, dew >= 40 & dew < 50)
temp6 = subset(temp, dew >= 50)

N0 = length(temp0)
N1 = length(temp1)
N2 = length(temp2)
N3 = length(temp3)
N4 = length(temp4)
N5 = length(temp5)
```

```
N6 = length(temp6)

d = (N-1)*var(temp) - sum((c(N0, N1, N2, N3, N4, N5, N6) -
1)*(c(var(temp0), var(temp1), var(temp2), var(temp3), var(temp4),
var(temp5), var(temp6))))
d
```

Output:

```
[1] 51024.52
```

C – Code to stratify dew point every 20˚F:

```
# creating strata from dew population (every 20)
df = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")
temp <- df[1:720, 5]
dew <- df[1:720, 9]
humidity <- df[1:720, 10]

N = 720
n = 84

temp0 = subset(temp, dew < 20)
temp1 = subset(temp, dew >= 20 & dew < 40)
temp2 = subset(temp, dew >= 40)

N0 = length(temp0)
N1 = length(temp1)
N2 = length(temp2)

d = (N-1)*var(temp) - sum((c(N0, N1, N2) - 1)*(c(var(temp0), var(temp1),
var(temp2))))
> d
[1] 41389.77
```

```
# stratified random sample w/ n = 84 within each strata
nh = 84/7 (equal allocation)

> dew1 = c()
> dew2 = c()
> dew3 = c()
```

```
> dew4 = c()
> dew5 = c()
> dew6 = c()
> dew7 = c()
> temp1 = c()
> temp2 = c()
> temp3 = c()
> temp4 = c()
> temp5 = c()
> temp6 = c()
> temp7 = c()
> N = 720

> for(i in 1:N){
+   if (dew[i] < 0) {
+     dew1 = append(dew1, dew[i])
+     temp1 = append(temp1, temp[i])
+   }
+   else if (dew[i] < 10) {
+     dew2 = append(dew2, dew[i])
+     temp2 = append(temp2, temp[i])
+   }
+   else if (dew[i] >= 10 & dew[i] < 20) {
+     dew3 = append(dew3, dew[i])
+     temp3 = append(temp3, temp[i])
+   }
+   else if (dew[i] >= 20 & dew[i] < 30) {
+     dew4 = append(dew4, dew[i])
+     temp4 = append(temp4, temp[i])
+   }
+   else if (dew[i] >= 30 & dew[i] < 40) {
+     dew5 = append(dew5, dew[i])
+     temp5 = append(temp5, temp[i])
+   }
+   else if (dew[i] >= 40 & dew[i] < 50) {
+     dew6 = append(dew6, dew[i])
+     temp6 = append(temp6, temp[i])
+   }
+   else if (dew[i] >= 50) {
+     dew7 = append(dew7, dew[i])
+     temp7 = append(temp7, temp[i])
+   }
+
+ }

> N1 = length(dew1)
> N1
[1] 17
> N2 = length(dew2)
> N2
[1] 49
> N3 = length(dew3)
```

```
> N3
[1] 133
> N4 = length(dew4)
> N4
[1] 236
> N5 = length(dew5)
> N5
[1] 167
> N6 = length(dew6)
> N6
[1] 92
> N7 = length(dew7)
> N7
[1] 26

> nh = 84/7
> s1 = sample(1:N1, nh, replace=F)
> s2 = sample(1:N2, nh, replace=F)
> s3 = sample(1:N3, nh, replace=F)
> s4 = sample(1:N4, nh, replace=F)
> s5 = sample(1:N5, nh, replace=F)
> s6 = sample(1:N6, nh, replace=F)
> s7 = sample(1:N7, nh, replace=F)

> temp1[s1]
 [1] 14.7 19.5 13.5 19.9 17.1 21.8 10.9 16.9 14.8 12.5 16.4 14.3
> temp2[s2]
 [1] 28.7 24.0 23.4 27.7 18.5 29.5 23.0 24.8 22.8 43.5 26.4 26.2
> temp3[s3]
 [1] 38.7 34.3 32.1 37.4 33.3 31.0 33.5 40.1 40.9 36.4 28.1 19.7
> temp4[s4]
 [1] 37.3 42.2 46.8 29.3 29.5 31.0 30.8 48.2 37.4 41.2 39.9 46.8
> temp5[s5]
 [1] 41.1 48.1 47.2 43.3 46.1 44.2 46.6 43.6 45.2 42.4 43.9 42.2
> temp6[s6]
 [1] 44.8 49.4 62.0 55.3 49.2 50.7 48.9 42.9 49.6 50.7 50.2 56.6
> temp7[s7]
 [1] 55.3 68.8 66.4 61.1 56.0 63.9 56.2 58.0 62.1 52.9 52.5 56.9
```

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

The estimated mean temperature is 39.958˚F with an estimated variance of 0.568. The 95% confidence interval is (38.383, 41.532) with a width of 3.149.

Output:

```
> y_bar_st = (sum((c(N1, N2, N3, N4, N5, N6, N7)* c(mean(temp1[s1]),
mean(temp2[s2]), mean(temp3[s3]), mean(temp4[s4]), mean(temp5[s5]),
mean(temp6[s6]), mean(temp7[s7])))))/N
> y_bar_st
```

```
[1] 39.95752

> var_bar = sum ( (c(N1, N2, N3, N4, N5, N6,N7)/N)^2 * ( (c(N1, N2,
N3, N4, N5, N6, N7)-nh) / c(N1, N2, N3, N4, N5, N6,N7) ) * (
c(var(temp1[s1]), var(temp2[s2]), var(temp3[s3]), var(temp4[s4]),
var(temp5[s5]), var(temp6[s6]),var(temp7[s7]) )/nh ) )
> var_bar
[1] 0.5676744

> numer = ( N1*(N1-nh)*var(temp1[s1])*(1/nh) +
N2*(N2-nh)*var(temp2[s2])*(1/nh) + N3*(N3-nh)*var(temp3[s3])*(1/nh)
+ N4*(N4-nh)*var(temp4[s4])*(1/nh) +
N5*(N5-nh)*var(temp5[s5])*(1/nh) + N6*(N6-nh)*var(temp6[s6])*(1/nh)
+ N7*(N7-nh)*var(temp7[s7])*(1/nh) )^2
> denom = (N6*(N6-nh)*(var(temp6[s6])/nh))^2/(nh-1) +
(N1*(N1-nh)*(var(temp1[s1])/nh))^2/(nh-1) +
(N2*(N2-nh)*(var(temp2[s2])/nh))^2/(nh-1) +
(N3*(N3-nh)*(var(temp3[s3])/nh))^2/(nh-1) +
(N4*(N4-nh)*(var(temp4[s4])/nh))^2/(nh-1) +
(N5*(N5-nh)*(var(temp5[s5])/nh))^2/(nh-1) +
(N7*(N7-nh)*(var(temp7[s7])/nh))^2/(nh-1)

> d = numer/denom
> d
[1] 19.44279

> y_bar_st + qt(c(0.025, 0.975), d)*sqrt(var_bar)
[1] 38.38298 41.53207
```

3. Take a stratified random sample with size *n* chosen in Report 2 with proportional allocation. ($n_h = \frac{n N_h}{N}$)

Output:

```
> #Proportional allocation
> n1 = round(n * N1 / N)
> n2 = round(n * N2 / N)
> n3 = round(n * N3 / N)
> n4 = round(n * N4 / N)
> n5 = round(n * N5 / N)
> n6 = round(n * N6 / N)
> n7 = round(n * N7 / N)

> n1
[1] 2
> n2
[1] 6
> n3
```

```
[1] 16
> n4
[1] 28
> n5
[1] 19
> n6
[1] 11
> n7
[1] 3

> s1 = sample(1:N1, n1, replace=F)
> s2 = sample(1:N2, n2, replace=F)
> s3 = sample(1:N3, n3, replace=F)
> s4 = sample(1:N4, n4, replace=F)
> s5 = sample(1:N5, n5, replace=F)
> s6 = sample(1:N6, n6, replace=F)
> s7 = sample(1:N7, n7, replace=F)

> temp1[s1]
[1] 13.8 14.7
> temp2[s2]
[1] 18.3 21.8 25.1 26.0 27.6 30.8
> temp3[s3]
 [1] 31.0 36.8 35.9 33.5 30.4 30.3 34.7 39.9 33.5 25.9 33.1 28.6
37.4 38.5 22.6 32.8
> temp4[s4]
 [1] 33.6 38.5 33.0 34.1 42.4 39.4 48.2 44.9 34.1 34.7 33.1 35.0
40.0 41.3 35.8 36.3 45.6 34.0 35.9 36.0 37.1 39.6 36.5 34.8 38.5
34.9 46.2 32.6
> temp5[s5]
 [1] 36.3 53.0 49.2 43.9 45.2 38.2 41.5 51.1 47.1 42.2 43.9 37.0
45.5 43.7 47.6 55.6 39.0 47.9 38.3
> temp6[s6]
 [1] 49.4 44.7 49.3 50.3 53.1 52.1 49.5 45.4 49.7 58.2 55.8
> temp7[s7]
[1] 54.8 52.9 62.2
```

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

The estimated true average temperature is 39.308℉ with estimated variance of 0.016. The 95% confidence interval is (39.055, 39.560) with a width of 0.505.

Output:

```
> y_bar_st = (sum((c(N1, N2, N3, N4, N5, N6, N7)* c(mean(temp1[s1]),
mean(temp2[s2]), mean(temp3[s3]), mean(temp4[s4]), mean(temp5[s5]),
mean(temp6[s6]), mean(temp7[s7])))))/N
> y_bar_st
[1] 39.3076
```

```
> var_bar = sum ( (c(N1, N2, N3, N4, N5, N6, N7)/N)^2 * (c(N1-n1,
N2-n2, N3-n3, N4-n4, N5-n5, N6-n6, N7-n7) / c(N1, N2, N3, N4, N5,
N6,N7) ) * c(var(temp1[s1])/n1, var(temp2[s2]/n2),
var(temp3[s3]/n3), var(temp4[s4]/n4), var(temp5[s5]/n5),
var(temp6[s6]/n6),var(temp7[s7]/n7) ) )
> var_bar
[1] 0.0160292

> numer = ( N1*(N1-n1)*var(temp1[s1])*(1/n1) +
N2*(N2-n2)*var(temp2[s2])*(1/n2) + N3*(N3-n3)*var(temp3[s3])*(1/n3)
+ N4*(N4-n4)*var(temp4[s4])*(1/n4) +
N5*(N5-n5)*var(temp5[s5])*(1/n5) + N6*(N6-n6)*var(temp6[s6])*(1/n6)
+ N7*(N7-n7)*var(temp7[s7])*(1/n7) )^2
> denom = (N6*(N6-n6)*(var(temp6[s6])/n6))^2/(nh-1) +
(N1*(N1-n1)*(var(temp1[s1])/n1))^2/(nh-1) +
(N2*(N2-n2)*(var(temp2[s2])/n2))^2/(n2-1) +
(N3*(N3-n3)*(var(temp3[s3])/n3))^2/(n3-1) +
(N4*(N4-n4)*(var(temp4[s4])/n4))^2/(n4-1) +
(N5*(N5-n5)*(var(temp5[s5])/n5))^2/(n5-1) +
(N7*(N7-n7)*(var(temp7[s7])/n7))^2/(n7-1)

> d = numer/denom
> d
[1] 71.98488

> y_bar_st + qt(c(0.025, 0.975), d)*sqrt(var_bar)
[1] 39.05522 39.55999
```

5. Take a stratified random sample with size *n* chosen in Report 2 with optimum allocation. ($n_h = \frac{n N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}$)

```
# stratified random sample w/ n = 84 within each strata  (optimum
allocation)

> bottom = N1*sqrt(var(temp1)) + N2*sqrt(var(temp2)) +
N3*sqrt(var(temp3)) + N4*sqrt(var(temp4)) + N5*sqrt(var(temp5)) +
N6*sqrt(var(temp6)) + N7*sqrt(var(temp7))
> n1 = ((n * N1 * sqrt(var(temp1))) / bottom)
> n2 = ((n * N2 * sqrt(var(temp2))) / bottom)
> n3 = ((n * N3 * sqrt(var(temp3))) / bottom)
> n4 = ((n * N4 * sqrt(var(temp4))) / bottom)
> n5 = ((n * N5 * sqrt(var(temp5))) / bottom)
> n6 = ((n * N6 * sqrt(var(temp6))) / bottom)
> n7 = ((n * N7 * sqrt(var(temp7))) / bottom)

> n1+n2+n3+n4+n5+n6+n7
[1] 84
```

```
> n1
[1] 1.379753
> n2
[1] 5.905547
> n3
[1] 15.97122
> n4
[1] 26.75884
> n5
[1] 19.90378
> n6
[1] 11.32028
> n7
[1] 2.760586

> s1 = sample(1:N1, n1, replace=F)
> s2 = sample(1:N2, n2, replace=F)
> s3 = sample(1:N3, n3, replace=F)
> s4 = sample(1:N4, n4, replace=F)
> s5 = sample(1:N5, n5, replace=F)
> s6 = sample(1:N6, n6, replace=F)
> s7 = sample(1:N7, n7, replace=F)
> temp1[s1]
[1] 16.4
> temp2[s2]
[1] 30.2 26.2 31.0 24.8 20.3
> temp3[s3]
 [1] 28.4 37.1 37.4 32.4 30.5 31.3 28.9 37.7 31.8 29.1 35.6 33.5
32.9 33.4 32.7
> temp4[s4]
 [1] 38.4 40.6 41.2 35.6 40.6 30.9 35.6 36.3 39.4 34.1 40.3 32.4
35.5 46.6 37.0 36.3 42.2 36.7 39.1 36.3 37.4 33.0 39.6 35.5 33.2
36.2
> temp5[s5]
 [1] 47.2 41.3 38.7 38.3 38.2 47.3 40.9 39.9 41.0 43.6 40.4 43.9
38.2 44.5 41.5 41.7 46.1 58.2 45.7
> temp6[s6]
 [1] 44.7 49.6 54.9 43.0 61.3 45.4 51.3 60.3 51.8 58.2 50.0
> temp7[s7]
[1] 54.8 55.3
```

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

The estimated true average temperature is 39.0705 with estimated variance of 0.012. The 95% confidence interval is (38.852, 39.289) with a width of 0.437.

```
> y_bar_st = (sum((c(N1, N2, N3, N4, N5, N6, N7)* c(mean(temp1[s1]),
mean(temp2[s2]), mean(temp3[s3]), mean(temp4[s4]), mean(temp5[s5]),
```

```
mean(temp6[s6]), mean(temp7[s7])))))/N
> y_bar_st
[1] 39.0705

> #Note: variance for first strata is 0 because n1 = 1, so the
variance would be 0 since it is just a sample of one.
> var_bar = sum ( (c(N1, N2, N3, N4, N5, N6, N7)/N)^2 * (c(N1-n1,
N2-n2, N3-n3, N4-n4, N5-n5, N6-n6, N7-n7) / c(N1, N2, N3, N4, N5,
N6,N7) ) * c(0/n1, var(temp2[s2]/n2), var(temp3[s3]/n3),
var(temp4[s4]/n4), var(temp5[s5]/n5),
var(temp6[s6]/n6),var(temp7[s7]/n7) ) )
> var_bar
[1] 0.0119485

> #Note: variance for first strata is 0 because n1 = 1, so the
variance would be 0 since it is just a sample of one.
> numer = ( N1*(N1-n1)*0*(1/n1) + N2*(N2-n2)*var(temp2[s2])*(1/n2) +
N3*(N3-n3)*var(temp3[s3])*(1/n3) + N4*(N4-n4)*var(temp4[s4])*(1/n4)
+ N5*(N5-n5)*var(temp5[s5])*(1/n5) +
N6*(N6-n6)*var(temp6[s6])*(1/n6) + N7*(N7-n7)*var(temp7[s7])*(1/n7)
)^2
> denom = (N6*(N6-n6)*(var(temp6[s6])/n6))^2/(n6-1) +
(N1*(N1-n1)*(0/n1))^2/(n1-1) +
(N2*(N2-n2)*(var(temp2[s2])/n2))^2/(n2-1) +
(N3*(N3-n3)*(var(temp3[s3])/n3))^2/(n3-1) +
(N4*(N4-n4)*(var(temp4[s4])/n4))^2/(n4-1) +
(N5*(N5-n5)*(var(temp5[s5])/n5))^2/(n5-1) +
(N7*(N7-n7)*(var(temp7[s7])/n7))^2/(n7-1)

> d = numer/denom
> d
[1] 60.31275

> y_bar_st + qt(c(0.025, 0.975), d)*sqrt(var_bar)
[1] 38.85187 39.28913
```

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

Width of CI with estimator using **equal allocation**: 41.53207 - 38.38298= 3.14909

Width of CI with estimator using **proportional allocation**: 39.55999 - 39.05522= 0.50477

Width of CI with estimator using **optimum allocation**: 39.28913 - 38.85187= 0.43726

Thus, the best estimator of true average winter temperature in College Park is the

one using stratified sampling with optimum allocation since its CI is the narrowest.

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

**All code and output are listed at each step above.