Exercise 1: Consider the data table from problem 9 in the last homework.

a. Take a SRS of size $n = 8$ from the top three rows, then another sample of size $n = 10$ from the bottom four rows. Then use stratified sampling estimator to get an estimate of the total over the entire region.

Solution:
Code:

```
x <- c(1, 3, 4, 5, 3, 3, 0, 7, 5, 0, 1, 4, 1, 4,
         2, 3, 4, 3, 4, 4, 3, 3, 0, 1, 2, 1, 5, 5,
         3, 3, 11, 9, 9, 15, 10, 6, 6, 12, 12, 12,
         18, 14, 13, 7, 10, 13, 3,14,11,11,15,11,
         13,12,12,13,19,10,11,14,13,6,9,7,15,14,
         9,16,13,12)

N <- c(length(x[1:30]),length(x[31:length(x)]))
n <- (8, 10)

strata_1 = sample(x[1:30],size = 8)
strata_2 = sample(x[31:length(x)],size = 10)

Stratified_Total = N[1]*mean(strata_1) + N[2]*mean(strata_2)
[1] 520.75

s_squared <- c(var(strata_1), var(strata_2))
[1] 0.5535714 15.1222222

Stratified_Total_std = sqrt(sum(N*(N - n)*(s_squared/n)))
[1] 43.132

CI <- c(Stratified_Total+2*Stratified_Total_std , Stratified_Total -2*St
[1] 607.0132 434.4868
```

b. How does the confidence interval in *a* compare to the one in the last homework? In particular, how wide are the intervals? Why do you think this occurred?

Solution:
Given that the standard error I from the simple random sample in problem 9 is actually smaller than the stratified sample above, the confidence interval was actually smaller before, without stratification. I think this can possibly be attributed to a poor grouping of strata that don't actually minimize the variance.

Exercise 2: I want to know the average contamination level in a plot of ground. Just by looking at the area, I see that part has discolored soil and damaged plats. I'll put this area into one stratum and the rest of the area into another stratum.
I collect $n = 3$ measurements in each stratum:
Stratum One (appears contaminated): $Area = 5$ Hectares, $Measurements = [50, 100, 75]$ ppt
Stratum Two (appears uncontaminated): $Area = 45$ Hectares, $Measurements = [1, 3, 2]$ ppt

1. Find a 95 percent confidence interval for the true mean contamination level.

Solution:
Code:

```
 strata_1 <- c(50, 100 ,75)
 strata_2 <- c(1,3,2)
N <-c(5, 45)
n <- c(3, 3)

W = N/sum(N)
 [1] 0.1 0.9

 Stratified_Mean = W[1]*mean(strata_1) + W[2]*mean(strata_2)
 [1] 9.3

 s_squared <- c(var(strata_1), var(strata_2))
 [1] 625 1
 Stratified_Mean_std = sqrt(sum(W^2*((N - n)/N)*(s_squared/n)))
 [1] 1.041793

CI <- c(Stratified_Mean+2*Stratified_Mean_std , Stratified_Mean -2*Stra
 [1] 11.383587 7.216413
```

2. Assuming the cost of collecting data is the same everywhere, use the variances from the study to plan next year's study, with the goal of getting an optimal allocation with margin of 1ppt.

Solution:
With constant cost, the equation for computing optimal allocation simplifies to,

$$w_i = \frac{N_i \sigma_i}{\sum_{j=i}^{K} N_j \sigma_j}$$

Code:

```
s_squared
[1] 625    1

Optimal_Allocation =
    (N*sqrt(s_squared))/sum((N*sqrt(s_squared)))
[1] 0.7352941 0.2647059

> NumberOfSamples =
    (sum(N^2*s_squared/Optimal_Allocation))/
        (sum(N)^2*(1/4)+sum(N*s_squared))
[1] 7.615283
```

Exercise 3: We wish to find the proportion of people who support a candidate in an election. We could use a SRS over the entire city, but instead we know one area ($N_1 = 12000$ voters) where our candidate isn't very popular(less than 30 percent is our guess) and another area ($N_2 = 24000$ voters) where we think or candidate will be more popular (over 50 percent, as a guess). We perform a SRS in stratum and get the following:
Stratum One: $N_1 = 12000$, $p_1 = .25$, $n_1 = 200$.
Stratum Two: $N_2 = 24000$, $p_2 = .65$, $n_2 = 400.4$.

a. Find a 95 percent confidence interval for the true proportion over the entire city.

Solution:
Code:

```
N <- c(12000,24000)
p <- c(.25,.65)
n <- c(200, 400.4)
W = N/sum(N)
    [1] 0.3333333 0.6666667

Stratified_Proportion = sum(W*p)
    [1] 0.5166667

Stratified_Proportion_std =
    sqrt(sum(W^2*((N - n)/N)*((p*(1-p))/(n-1))))
    [1] 0.01875845

CI = c(Stratified_Proportion+2*Stratified_Proportion_std,
       Stratified_Proportion - 2*Stratified_Proportion_std)
    [1] 0.5541836 0.4791498
```

b. Is this proportional allocation? Why or why not? If so, why did I use proportional allocation? If not, why did I choose the sample sizes I did?

Solution:
This is an example of proportional allocation. We can see that the sample sizes were chosen as a proportionally with respect to the size of the stratum. Proportional allocation gives an unbiased estimator when the true proportion or even an estimate isn't had to compute the optimal allocation.

c. Was stratification a good idea in this case? Why or why not?

Solution:
Computing the estimated proportion as one SRS we can find out if stratifying reduced the SE in any useful way.
Code:

```
SRSProportion = sum(p*n)/sum(n)
[1] 0.5167555

SRSProportion_std = sqrt((sum(N) - sum(n))/sum(N)*
            (SRSProportion*(1 - SRSProportion))/(sum(n)-1))
[1] 0.02024024

> CI = c(SRSProportion+2*SRSProportion_std,
            SRSProportion - 2*SRSProportion_std)
[1] 0.557236 0.476275
```

From the code we can see that the SE was reduced .2%, when we stratified. In general we should see a reduction in the SE when stratum proportions $\hat{p}_i$ are more ideal than .50. Elections have been won on margins less than .04% so I would say stratification is a good idea in this case.

Exercise 4: For the results in problem 3, if we had known the proportions were around .25 and .65, what would be the optimal allocation of the sample between the two strata? If we wanted a bound of ±.05, what sample size should I take. (Assuming constant cost)

Solution:
Code:

```
Optimal_Allocation = N*sqrt(p*(1-p))/sum(N*sqrt(p*(1-p)))
    [1] 0.3122046 0.6877954
```

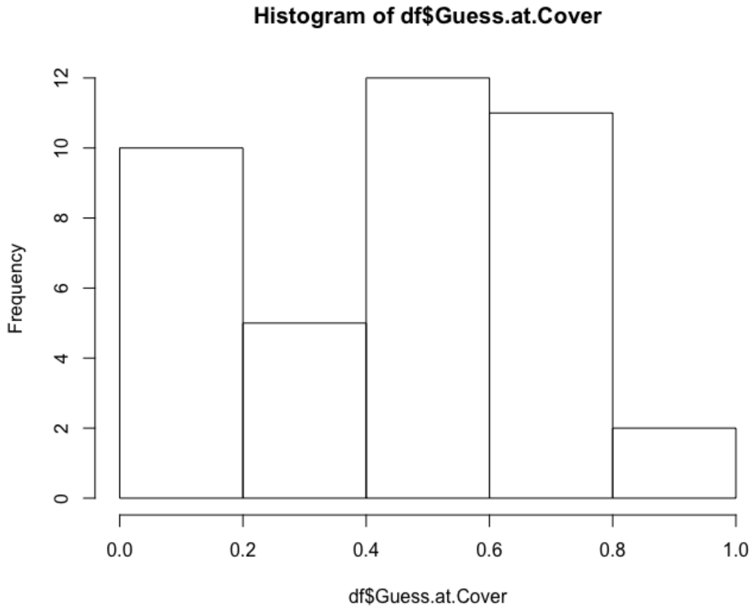
```
Total_Samples = sum(N^2*(p*(1-p))/Optimal_Allocation)/
                (sum(N)^2*((.05)^2/4)+sum(N*(p*(1-p))))
    [1] 338.7552
```

Exercise 5: We get to see an entire population once again. Yep, that happens a lot in practice... We want to take a total sample of size $n = 9$ from the population of $N = 40$ equal-area plots. We want to estimate the total grass cover in the region. Measuring that is not easy, but it is easy to quickly look over the plots and 'guess' at the grass cover. below is the real population(which we magically see) along with the 'guess' at grass cover, which we do know for all of the plots.
In other words, start by pretending that you can't see the second column. the use then use these values to divide the plot into strata. Finally, you will take a sample in each of the strata, where you can write down the value of the second column for the plots you sample.

a. In reality, you only know the guess until you take your sample. Use the guesses to stratify the population into three strata. Tell me how you decided upon a sample size for each. Then use a randomization method to select three plots from each stratum. Then write down the 'true cover' for those plots.

   Solution:
   To group the plots up into strata I simply plotted a histogram with the minimal number of bins possible. Doing so gave me this histogram,

Figure 1: Histogram of the guess cover.



Looking at the histogram I decided to group the strata with the following criteria,
Code:

```
Strata_1 = subset(df, Guess.at.Cover <= .4)
Strata_2 = subset(df, Guess.at.Cover > .4 & Guess.at.Cover <= .8 )
Strata_3 = subset(df, Guess.at.Cover > .8 )

sample(Strata_1$True.Cover, 3 )
    [1] 0.10 0.20 0.15
sample(Strata_2$True.Cover, 3 )
    [1] 0.5 0.7 0.6
sample(Strata_3$True.Cover, 2)
    [1] 0.75 0.75
```

b. With the data from *a*. find the 95 percent confidence interval for the mean cover.

Solution:
Code:

```
N <- c(length(Strata_1$Plot), length(Strata_2$Plot), length(Strata_3$Pl
    [1] 15 23  2
```

```
W = N/sum(N)
    [1] 0.375 0.575 0.050

## This sample was chosen with sample(), I just forgot to assign it.
Sampled_Strata_1 = c(0.10, 0.20 ,0.15)
Sampled_Strata_2 = c(0.5, 0.7, 0.6)
Sampled_Strata_3 = c(0.75, 0.75)

##  Computing the mean of each strata
Mean_Strata_1 = mean(Sampled_Strata_1)
Mean_Strata_2 = mean(Sampled_Strata_2)
Mean_Strata_3 = mean(Sampled_Strata_3)
Mean_Strata = c(Mean_Strata_1, Mean_Strata_2, Mean_Strata_3)

Stratified_Mean = sum(W*Mean_Strata)
    [1] 0.43875

## Computing S^2 for each strata
Var_Strata_3 = var(Sampled_Strata_3)
Var_Strata_2 = var(Sampled_Strata_2)
Var_Strata_1 = var(Sampled_Strata_1)
Var_Strata = c(Var_Strata_1, Var_Strata_2, Var_Strata_3)


SE = sqrt(sum(W^2*((N - c(3,3,2))/N)*(Var_Strata/c(3,3,2))))
    [1] 0.03243583

CI = c(Stratified_Mean + 2*SE, Stratified_Mean - 2*SE)
    [1] 0.5036217 0.3738783
```

Looking at the confidence interval, we can see that we do contain the true mean cover, and our estimate is also within 1 standard error.