Exercise 1: I want to survey a group of students on how much time they are spending on homework and studying. I have a frame of students that I can contact, along with their age, gender, major and year in college. Do you think it would be worth stratifying on one of these variables? Why or why not? If you were to use a variable to make strata, which would you pick and why?

Solution:
I think it is absolutely worth stratifying on one if not two of these variables. Firstly, considering a students major, it's easy to see that there is significant disparity when it comes to study time. Furthermore there is also likely significant disparity in study time when comparing upper and lower class men, as classes tend to scale up in difficulty/required study time. Therefore a SRS that contains lower and upper class men, rigorous and non-rigorous majors will almost assuredly have a higher variance with respect to study time than if you had stratified.

Exercise 2: In the Alaska Department of Fish and Game paper (included with this HW) they use either stratification or post-stratification. Which one did they use? What were the strata? Why did they do this?

Solution:
In the 2002 ADF&G survey for Chinook Salmon we can see that the primary sampling method was regular stratification with some secondary use of post-stratification. In the section of the paper it is stated that the area of interest was split into three separate stratum, the lower, middle, and upper portions of the Kuskokwim River. The paper states that the reason for dividing the area into these stratum is because of ”differing proportions in gear type usage”. Beyond that each SRS in the stratum was designed as an ”opportunistic” sample i.e. samples were taken across time with a variety of gear with the assumption that samples would still be unbiased and independent. It is also stated, later in the sample design section that the SRS in each stratum would be ”post stratify(ed) by time and gear”. Under an ”opportunistic” sampling scheme this, further post-stratification seems necessary to account for variance across sampling time and gear. The details of the post-stratification is defined in the second paragraph of the ”Data Processing, Analysis, and Reporting” section.

Exercise 3: We really wish to estimate the total expenditures for fuel in a city with $N = 80000$ households. We use random phone dialing to find a SRS of $n = 400$ households. However, we can divide the city into three strata, which we think will include houses with low fuel expense (first stratum), medium fuel expense (second stratum) and high fuel expense (third stratum) based on the typical temperatures and ages of houses in the three regions. We know the size of each stratum ($N1 = 20000$, $N2 = 30000$, $N3 = 30000$). However, we can't take a SRS of each stratum since phone numbers are only roughly related to address (and we are using random phone dialing).

However, we can sort the $n = 400$ sampled households into three samples, one from each stratum. We get the following:

* Stratum One: $\overline{x}_1 = \$2500.00$, $s_1 = \$500.00$, $n_1 = 120$.
* Stratum Two: $\overline{x}_2 = \$4000.00$, $s_2 = \$750.00$, $n_2 = 150$.
* Stratum Three: $\overline{x}_3 = \$5500.00$, $s_3 = \$750.00$, $n_3 = 130$.

a. What type of sampling is this?

Solution:
This is an example of post-stratification and it's when you divide the large SRS of size 400 into strata after you collect the samples.

b. Why is this better than just ignoring the strata and considering this to be an old, boring SRS of size $n = 400$ which we did in the beginning of the course?

Solution:
This is better for the same reason a regular stratified sampling schema is better. Recall the formula for calculating sample variance,

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}.$$

Note that having large disparity between samples, by definition will give us more variance. By grouping the data into strata we are able to reduce the variance in our estimator for the mean or total.

c. Find a 95 percent confidence interval for the true mean fuel expenditure in the city. What is a 95 percent confidence interval.

Solution:
Code:

```
> Strata_means <- c(2500, 4000, 5500)
> Strata_SSquared <- c(500, 750, 750)
> Strata_n <- c(120, 150, 130)
> Strata_N <- c(20000,30000,30000)
> Strata_Proportion = Strata_n/400

> Mean_estimator = sum(Strata_Proportion*Strata_means)
[1] 4037.5

> Margin_error = 2*sqrt(sum(Strata_Proportion^2
                          *((Strata_N - Strata_n)/Strata_n)
                          *(Strata_SSquared/Strata_n)))

> CI95 <- c(Mean_estimator + Margin_error,
            Mean_estimator - Margin_error)
[1] 4074.49 4000.51
```

Finally we get $\hat{\mu} = 4037.5$ and a 95 percent confidence interval of (4074.49, 4000.51). A 95 percent confidence interval means that there is 95% chance the true mean is contained in the interval.

d. Suppose that we complete the study and decide, before beginning the analysis, to just analyze it without sorting into strata. Would this be valid? Why or why not?

Solution:
This would be a valid technique. Usually when you have to use post-stratification the sampling schema is already set up for a 'one big' SRS analysis. However the point of post-stratification is that we have information about how the data can be stratified and ignoring that information means leaving accuracy on the table.

Exercise 4.: We want to estimate the concentration of available nitrogen in the soils of a region. Cold, wet soils generally have more available nitrogen, but also soils with nitrogen fixing (alder) plants or areas showing high productivity might have higher soil nitrogen. We think we can very easily classify plots of ground into either low or high nitrogen plots just by looking at them, but the actual soil sampling and analysis is expensive.

To lower cost, we'll do the following:

(1) divide the region into $N = 20000$ reasonably-sized plots.
(2) take a SRS of size $m = 500$ plots which we will visit and rapidly classify into either high N stratum or low N stratum (actually this would probably be done as a systematic sample, not an SRS, which we'll see later).
(3) We find that $m_1 = 300$ of these plots are classified as low nitrogen and $m_2 = 200$ plots as high nitrogen.
(4) Now we take an SRS of size $n_1 = 30$ from the low nitrogen (we hope) plots and, independently, $n_2 = 50$ from the high nitrogen plots. We get the following:

Classified as low nitrogen: $\overline{x}_1 = 30ppm$, $s_1 = 10ppm$
Classified as high nitrogen: $\overline{x}_2 = 40ppm$, $s_2 = 15ppm$.

a. Does it appear that the stratification will help us much? If we decide it doesn't, can we just pretend we took a SRS of size 500 and ignore the stratification?

b. Find a 95 percent confidence interval for the average nitrogen concentration.

Solution:
Code:

```
> Strata_means <- c(30, 40)
> Strata_SSquared <-c(10, 15)
> Strata_n <- c(30, 50)
> Strata_m <- c(300, 200)
> Strata_Proportion = Strata_m/sum(Strata_m)

> Mean_estimator = sum(Strata_Proportion*Strata_means)
[1] 34
> Margin_error = 2*sqrt(sum(Strata_Proportion^2
                          *((Strata_m - Strata_n)/Strata_n)
                          *(Strata_SSquared/Strata_n)))

[1] 2.212691
```

```
> CI95 <- c(Mean_estimator + Margin_error ,
            Mean_estimator - Margin_error)
[1] 36.21269 31.78731
```

I do think that had we taken a SRS of size 500 we would have a higher variance than just using stratification. Beyond that, because of how the samples were stratified before they were collected we cannot redo the analysis as a large SRS, since the selection of the samples is no longer random.