STA 4222 **Homework 4 Assignment** **Due at 2pm on April 8th, 2020**

1. (Ex. 4.15 on page 158) Jackson et al. (1987) compared the precision of systematic and stratified sampling for estimating the average concentration of lead and copper in the soil. The $1 - km^2$ area was divided into 100-m squares, and a soil sample was collected at each of the resulting 121 grid intersections. Summary statistics from this systematic sample are given below.

Element	n	Average $(mg\ kg^{-1})$	Range $(mg\ kg^{-1})$	Standard Deviation $(mg\ kg^{-1})$
Lead	121	127	22-942	146
Copper	121	35	15-90	16

The investigators also post-stratified the same region. Stratum A consisted of farmland away from roads, villages, and woodlands. Stratum B contained areas within 50m of roads, and was expected to have larger concentrations of lead. Stratum C contained the woodlands, which were also expected to have larger concentrations of lead because the foliage would capture airborne particles. The data on centration of lead and copper were not used in determining the strata. The data from the grid points falling in each stratum are in the following table:

Element	Stratum	n_h	Average $(mg\ kg^{-1})$	Range $(mg\ kg^{-1})$	Standard Deviation $(mg\ kg^{-1})$
Lead	A	82	71	22-201	28
Lead	B	31	259	36-942	232
Lead	C	8	189	88-308	79
Copper	A	82	28	15-68	9
Copper	B	31	50	22-90	18
Copper	C	8	45	31-69	15

(a) Calculate a 95% CI for the average concentration of lead in the area, using the systematic sample. (You may assume this sample behaves like an SRS.) Repeat for the average concentration of copper.

(b) Now use the post-stratified sample, and find 95% CIs for the average concentration of lead and copper. How do these compare with the CIs in (a)? Do you think that using stratification in future surveys would increase precision?

2. (Ex. 5.5 on page 209) A language school owner takes an SRS of 10 of the 72 Introductory Spanish classes offered by the school. Each student in each of the sampled classes is given a vocabulary test and is also asked whether he or she is planning a trip to a Spanish-speaking country in the next year. The data are in file *spanish.csv*.

(a) Estimate the total number of students planning a trip to a Spanish-speaking country in the next year, and given a 95% CI.

(b) Estimate the mean vocabulary test score for Introductory Spanish students in the language school, and give a 95% CI.

3. (Ex. 5.7 on page 209) The new candy Green Globules is being test-marketed in an area of upstate New York. The market research firm decided to sample 6 cities from the 45 cities in the area and then to sample supermarkets within cities, wanting to know the number of cases of Green Globules sold.

City	Number of Supermarkets	Number of Cases Sold
1	52	146,180,251,152,72,181,171,361,73,186
2	19	99,101,52,121
3	37	199,179,98,63,126,87,62
4	39	226,129,57,46,86,43,85,165
5	8	12,23
6	14	87,43,59

Obtain summary statistics for each cluster. Plot the data, and estimate the total number of cases sold, and the average number sold per supermarket, along with the standard errors of your estimates.