HOMEWORK THREE:

This first problem uses the data table from the last HW (used in problem 9).

1. (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 the stratified sampling estimator to get an estimate of the total over the entire region.

(b) How does the confidence interval in (a) compare to the one from problem 9 in the last HW? In particular, how wide are the intervals? Why do you think this occurred?

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 plants. 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 ppt, 100 ppt, 75 ppt.

Stratum Two (appears uncontaminated): Area = 45 hectares, Measurements = 1 ppt, 3 ppt, 2 ppt.

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

b) Assuming that the cost of collecting data is the same everywhere, use the variances from the study to plan next year?s study, with a goal of getting an optimal allocation with a margin of error of 1 ppt.

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 our candidate will be more popular (over 50 percent, as a guess). We perform a SRS (independent) in each stratum and get the following:
Stratum One : $N1 = 12000,\ p_1 = 0.25,\ n1 = 200.$
StratumTwo: $N2 = 24000,\ p_2 = 0.65,\ n2 = 400.4$

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

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

4. For the results in problem 3, if we had known the proportions were around 0.25 and 0.65, what would be the optimal allocation of the sample between the two strata? If we wanted a bound of plus or minus 0.05, what sample size should I take?

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 a 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 get to 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 (you only have the guesses you made while quickly examining all of the plots). The

use these values (third column) to divide the plots 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 (that's the 'real' data).

Plot	True Cover	Guess at Cover
1	0.40	0.4
2	0.40	0.3
3	0.60	0.5
4	0.65	0.7
5	0.85	0.8
6	0.00	0.1
7	0.15	0.3
8	0.65	0.6
9	0.50	0.5
10	0.35	0.3
11	0.60	0.8
12	0.75	0.9
13	0.50	0.6
14	0.60	0.6
15	0.50	0.5
16	0.50	0.6
17	0.55	0.7
18	0.10	0.2
19	0.50	0.5
20	0.70	0.8
21	0.10	0.2
22	0.75	0.7
23	0.50	0.6
24	0.75	1.0
25	0.70	0.7
26	0.10	0.2
27	0.65	0.7
28	0.10	0.0
29	0.45	0.5
30	0.20	0.1
31	0.10	0.2
32	0.85	0.8
33	0.70	0.7
34	0.15	0.2
35	0.10	0.1
36	0.45	0.4
37	0.75	0.8
38	0.40	0.5
39	0.25	0.2
40	0.50	0.6

(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 stratum. Then use a randomization method to select three plots from each stratum. Then write down the 'True Cover' for those plots.

(b) With the data from (a), find the 95 percent confidence interval for the mean cover. NOTE: The actual population mean is $\mu = 0.46$. Did your interval contain this value?