Assignment 2

Deadline to hand in: Nov. 10th in class

Q. 1 Consider a population of 6 students. Suppose we know the test scores of the students to be

Student	1	2	3	4	5	6
Score	66	59	70	83	82	71

(a) Find the mean $\bar{y}_U$ and variance S^2 of the population.

(b) How many SRS's of size 4 are possible?

(c) List the possible SRS's. For each, find the sample mean and $V(\bar{y})$.

(d) Now let stratum 1 consists of students 1-3, and stratum 2 consists of students 4-6. How many stratified random samples of size 4 are possible in which 2 students are selected from each stratum?

(e) List the possible stratified random samples. Which of the samples from (c) can not occur with the stratified design?

(f) Find $\bar{y}_{str}$ for each possible stratified sample. Find $V(\bar{y}_{str})$, and compare it with $V(\bar{y})$.

Q. 2 We want to estimate the proportion of UTSC students living in on-campus who prefer not to eat at campus dining halls. We stratified based on students' classification such as freshmen, sophomore, junior and senior. We obtained preliminary estimates for each stratum proportion p_h by a pilot study. Upper year students are harder to reach, so it costs more to collect data from them. In the following table, we summarized the information from the pilot study and cost estimates.

H Stratum	N_h Studnes	$\hat{p}_h$ Prefer not to eat in dinning hall	C_h $ per sample student
Freshman=1	6812	0.37	3.00
Sophomore=2	4586	0.52	4.50
Junior=3	2714	0.68	4.50
Senior=4	618	0.84	6.00

(1) Suppose we are planning to use a sample size of n=200.

(a) What sample sizes would we have under proportional allocation?

(b) What sample sizes would we have under Neyman allocation?

(c) What sample sizes would we have under optimal allocation?

(d) Suppose that we have \$700 to spend and c_0 = \$20. What sample size would we have under this cost constraint using the allocation strategy in (c).

(2) Now, suppose we are mainly concern with estimating the proportion of seniors who prefer not to eat in campus dining halls.

(a) If we make a margin of error of 0.05 for a 95% confidence interval, how many seniors do we need to select for our sample?

(b) Suppose that we can only afford to devote 50% of our variable cost budget ($680) on this objective. How would you relax the requirements for the sample size determination for seniors in (a)?

Q.3 The following data are from a stratified sample of faculty, using the areas biological sciences, physical sciences, social sciences, and humanities as the strta.

Stratum	Number of Faculty Members in Stratum	Number of Faculty members in Sample
Biological Sciences	102	7
Physical Sciences	310	19
Social Sciences	217	13
Humanities	178	11
Total	807	50

The frequency table for number of publications in the strata is given below.

	Number of Faculty Members			
Refereed Publications	Biological	Physical	Social	Humanities
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

(a) Estimate the total number of referred publications by faculty members in the college, and give the standard error.

(b) Did stratification increase precision in this example? Explain why you think it did or did not.

(c) Estimate the proportion of faculty with no referred publications, and give the standard error.

Q. 4 The data used in this question ,"golfsrs.csv" and the documentation for the dataset are posted at Blackboard. The data set contains data on a SRS of 120 golf courses. The sample was selected from the list of 16,883 US golf courses, obtained from the website www.golfcourse.com.

(a) Estimate the average greens fee to play 9 holes on a weekend.

(b) Consider the relationship between the greens fee to play 9 holes on a weekend and the back-tee yardage.

i. Create a plot for these two variables.

ii. Estimate the correlation coefficient for these two variables

iii. Estimate a regression model for predicting weekend greens fees for 9 holes on a weekend from the back-tee yardage

(c) Suppose you were asked to obtain a regression estimate for the mean greens fee to play 9 holes on a weekend for the all golf courses listed on the website, along with a 90% confidence interval. Do you have all of the information you need? If not, what information is missing?

(d) Based on what you know about regression estimation, information from (b), would you expect the regression estimator to provide a better estimate of the mean greens fee for 9 holes on a weekend than the sample mean would? Justify your answer. Compute estimated SE of mean green fee to play 9 holes on a weekend, and estimated SE of regression estimate of mean greens fee to play 9 holes on a weekend, and comment on them.

Q. 5 A certain statistician wants to estimate the average number of hotdogs eaten by apartment dwellers last night in a certain area. A simple random sample of 20 apartment complexes out of 124, then the number of hotdogs eaten last night by each person living in the apartment was recorded and summed over the whole apartment complex (they were a very cooperative group and all responded.) The data file, "hotdogs.txt", that contains three variables, complex ID, number of dwellers and number of hotdogs eaten, is available on the Blackboard.

(a) Explain why this is a cluster sample.

(b) Estimate the average number of hotdogs eaten by the apartment folk last night. Also provide the standard error.