1. Sample Size Allocation:
   Suppose we want to estimate the proportion of college students living in the residence halls who prefer not to eat in the dining halls. We have decided to stratify based on student classifications (freshman, sophomore, junior, senior). Using a pilot study, we obtained preliminary estimates for each stratum proportion, p_h. Also, since upper class students are harder to reach, it costs more to collect data from them. The following table summarizes information from the frame, the pilot study, and cost estimates:

H Stratum	N_h Students	$\hat{p}_h$ Prefer not to eat in dining hall	C_h \$ per sample student
Freshman=1	6,900	0.30	3.00
Sophomore=2	4,600	0.50	4.50
Junior=3	2,800	0.70	4.50
Senior=4	700	0.80	6.00

   a. Suppose we are planning to use a sample size of $n = 200$.
      i. What sample sizes would we have under proportional allocation?
      ii. What sample sizes would we have under Neyman allocation?
      iii. What sample sizes would we have under optimal allocation?
      iv. Suppose that we have \$700 to spend and $c_o = \$20$. What sample size would we have under this cost constraint using the allocation strategy in (a.iii)?

   b. Suppose we also want to make inferences for each stratum. That is, we are interested in the 4 subpopulations defined by the student classification variable.
      i. What concerns do you have with respect to this goal about the allocation of units using each of the methods in a. i-iii?
      ii. Suggest a compromise allocation of units that addresses both the population and subpopulation objectives. Ignore the cost constraints.

   c. Now suppose we are mainly concerned with estimating the proportion of seniors who prefer not to eat in the dining hall.
      i. If we want a margin of error of 0.05 for a 95% confidence interval, how many seniors do we need to select for our sample?
      ii. Suppose we want to only 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 (c.i)?

2. For this problem, use the attached files to answer the following questions.

   a. Take a stratified random sample of 150 players form the file **baseball.csv** using the proportional allocation with the different teams as strata. Read in the population data using the following SAS statement and use 20201029 as a seed number.

```
filename baseball "C:\stat311\baseball.csv";

data baseball;
infile baseball delimiter="," ;
input team $ leanguID $ player $ salary POS $ G GS InnOuts
      PO A E DP PB GB AB R H SecB ThiB HR RBI SB CS BB
      SO IBB HBP SH SF GIDP;
```

b. Estimate the population mean of **log of salary[=**$\ln(salary)$**]** using your stratified sample and provide 95% confidence interval.

c. Estimate the proportion of players of each position(POS) using your stratified sample and give a 95% confidence interval.

d. Examine the sample variances (s_h^2) of **log of salary[=**$\ln(salary)$**]** in each stratum. Do you think optimal allocation would be worthwhile for this problem?

e. Using the sample variances obtained in (d), determine the optimal allocation for a sample in which the cost is the same in each stratum and the total sample size 150.