This problem set covers material from Week 1, dates 9/9 – 9/12.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Canvas assignment. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A general rubric for homework problems appears on the final page of this assignment.

Monday 9/9

None!

Wednesday 9/11

1. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all the videos on YouTube so you use a random video picker to select 1000 videos, You find that 2% of these videos are cat videos. Identify each of the following in this example:
   (a) A variable
   (b) The sample statistic
   (c) The population parameter
   (d) A case

2. The registrar keeps an alphabetical list of all undergraduates, along with their email addresses. Suppose there are 2,770 undergraduates at Middlebury College this semester. Someone proposes to choose a number at random from 1 to 50, count that far down the list, taking that name and every 50th name after it for the sample.
   (a) Is this a probability sampling method?
   (b) Is this sampling method the same as simple random sampling?
   (c) Is there selection bias in this method of drawing a sample?
   (d) Now suppose that I want a sample size of around 50 undergraduates. Using methods described above, what would I have to modify in order to obtain such a sample size?

3. (From the practice problems) To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the

researchers assessing the patients when they were sick did not. No statistically discernible differences were observed in any measure of cold duration or severity between the four groups, and the placebo group had the shortest duration of symptoms.

(a) Was this an experiment or an observational study? Why?

(b) What are the explanatory and response variables in this study?

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

(e) Participants are ultimately able to choose whether to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

4. A common saying is "An apple a day keeps the doctor away", meaning that eating an apple every day keeps you healthy. Suppose you and a team of researchers want to know if eating an apple a day really does keep the doctor away.

   Devise an experiment that answers this research question, meets the three principles of experimental design, and has a treatment/control group. Pat yourself on the back if you can make your experiment blinded! **Also state what your response variable of interest is; otherwise, you might be not be addressing the research question!** *Note: there is no single correct answer/experiment! Have fun with this!*

Thursday 9/12

5. Time to implement the sampling methods we learned! **Keep track of your work on this problem; we will return to it in the next few classes.**

   We have a farmer who grows sunflowers for making sunflower oil. Her field is arranged in a grid pattern, with 12 rows and 12 columns as shown below. Water is important for crops, so irrigation ditches have been installed along the top and bottom of the field. It is expected that plants closer to a water source will perform better than those further away from water.

   The farmer would like to estimate the number of healthy plants in the field, along with a few other characteristics about the sunflowers. It would be unfeasible to conduct a census, so we should choose to sample a subset of the grid cells. **Suppose we'd like to sample $n = 12$ grid cells total.**

   (a) Using words (sentences or bulleted list are fine) and perhaps labeling the figure below, describe exactly how you would obtain a sample of 12 grid cells using **simple random sampling**. I should be able to read your work and know what to do without any questions! Think about how you will perform the random sampling. (E.g. If you use code, write down the code you use (no need to include a .Rmd file). If you use some physical objects, describe the objects and how you will use them.)

Then implement the method that you've written down, and either using the corresponding figure below or drawing your own 12x12 field, shade in the squares that correspond to your sample.

(b) Using words (sentences or bulleted list are fine) and perhaps labeling the figure below, describe exactly how you would obtain a sample of 12 grid cells using **stratified sampling** where the strata are *rows*. I should be able to read your work and know what to do without any questions!

Then implement the method that you've written down, and either using the corresponding figure below or drawing your own 12x12 field, shade in the squares that correspond to your sample.

(c) Using words (sentences or bulleted list are fine) and perhaps labeling the figure below, describe exactly how you would obtain a sample of 12 grid cells using **stratified sampling** where the strata are *columns*. I should be able to read your work and know what to do without any questions!

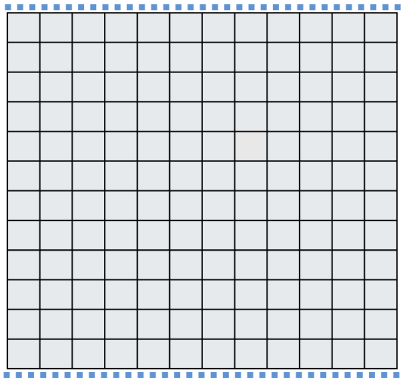
Then implement the method that you've written down, and either using the corresponding figure below or drawing your own 12x12 field, shade in the squares that correspond to your sample.

(d) Using words (sentences or bulleted list are fine) and perhaps labeling the figure below, describe exactly how you would obtain a sample of 12 grid cells using **cluster sampling** where we have 24 clusters total. I should be able to read your work and know what to do without any questions!

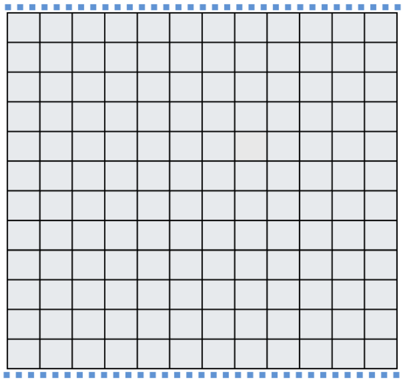
Then implement the method that you've written down, and either using the corresponding figure below or drawing your own 12x12 field, shade in the squares that correspond to your sample.

(e) Using words (sentences or bulleted list are fine) and perhaps labeling the figure below, describe exactly how you would obtain a sample of 12 grid cells using **multistage cluster sampling** where we have 24 clusters total. I should be able to read your work and know what to do without any questions!

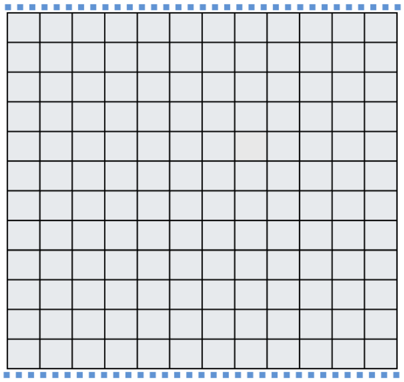
Then implement the method that you've written down, and either using the corresponding figure below or drawing your own 12x12 field, shade in the squares that correspond to your sample.



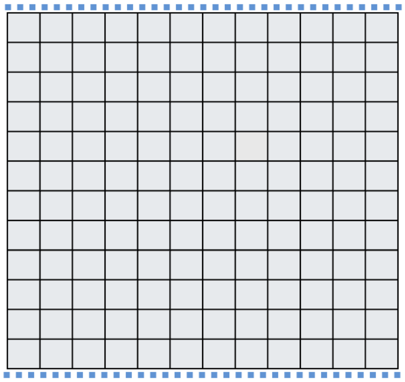
a) Simple random



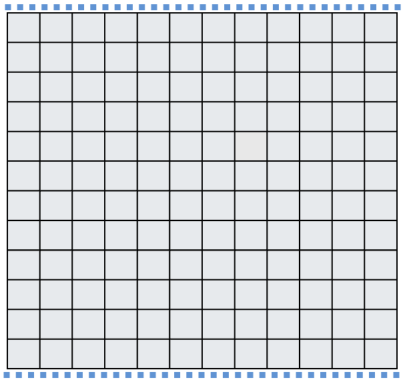
b) Stratified (rows)



c) Stratified (columns)



d) Cluster



e) Multistage cluster

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having lines of code that are necessary to solving the problem, as well as presenting the solution for the reader to easily see. It might also be worth adding comments to your code.