This problem set covers material from Week 6, dates 10/13 – 10/16.

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 10/13

1. Look to the live code associated with today's content.

   (a) Explain what each of lines 4-9 are doing.

   (b) In line 9, I used the function `sum()`. I claim that I could also somehow use the function `mean()`. Write for me code to replace line 9 that correctly utilizes the `mean()` function.

2. I have the following sample from a target population of interest: $\boldsymbol{x} = (4, 1, 2, 0, 2, 5)$.

   (a) Determine if each of the following is a valid bootstrap re-sample. If not, justify why not.

      i. $(4, 2, 1, 5, 8, 1)$
      ii. $(1, 1, 1, 1, 1, 1)$
      iii. $(1, 1,\ 2,\ 0,\ 4)$

   (b) Suppose the parameter of interest is the maximum value in the distribution. Would the bootstrap distribution of the sample maximum would yield a good approximation to the sampling distribution? Why or why not?

3. Suppose rather than measuring a single variable for each case, I observe pairs of observations (x_i, y_i) from every case, for $i = 1, \ldots, n$. For example, x_i could be the time it takes for person i to run a mile before participating in a training program, and y_i the time it takes for person i to run a mile after the program. I'd like to learn about the effect of the training program.

   (a) Either in words or "pseudo-code" (i.e. a mix of code and words), describe how I could obtain a bootstrap distribution of the average change in mile time. Please be clear and provide sufficient detail so I could implement your method.

   (b) If I want to demonstrate that the program is effective, how do I hope the bootstrap distribution looks like (i.e. tell me about center, spread, and maybe shape)?

   (c) How would you use your bootstrapping scheme to estimate the standard error of the sample mean?

Wednesday 10/15

4. Work on problems in associated `.qmd` document. Problems 2 and 3 will be graded as one problem, as will 4 and 5.

Thursday 10/16

5. People providing an organ for donation sometimes seek the help of a special medical consultant. These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients.

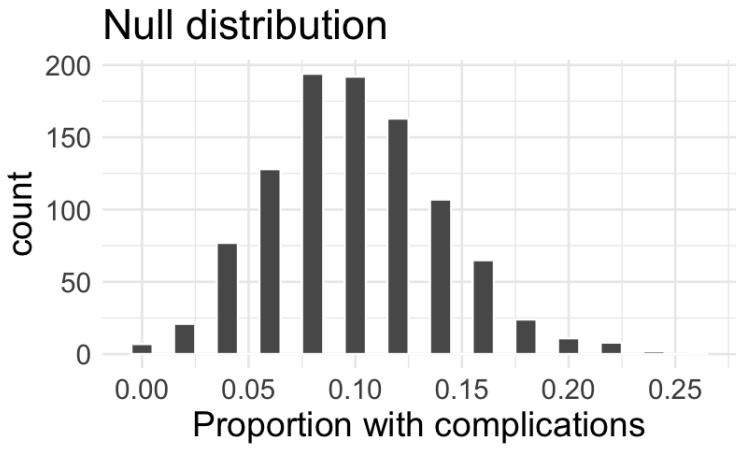
   One consultant tried to attract patients by noting that the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 4 complications in the 70 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

   (a) The consultant's claim is causal. Is it possible to assess the consultant's claim using the data?

   (b) Using proper statistical notation and defining quantities where necessary, state the hypotheses to test the consultant's claim.

   (c) In words, describe how you would conduct a simulation scheme to obtain a null distribution for the sample statistic. Also describe how you would use the null distribution to calculate the p-value. *Be as specific as possible.*

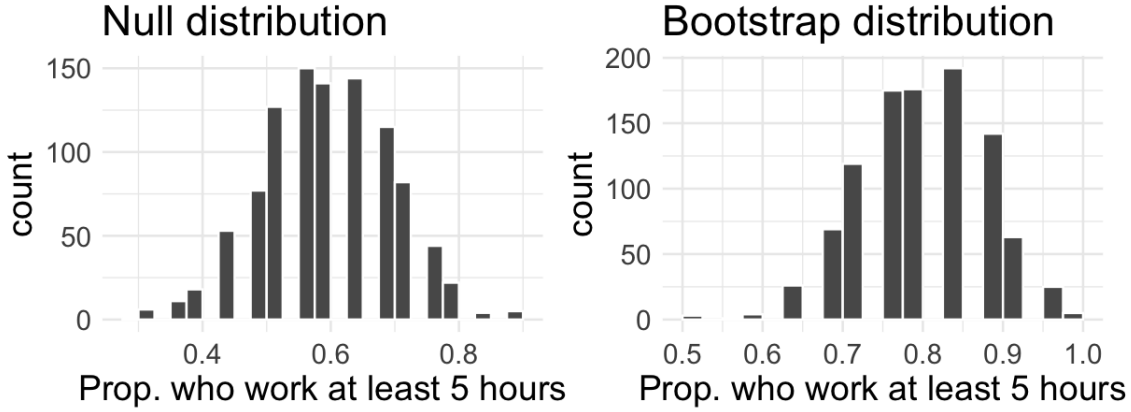
   (d) A histogram of the null distribution is shown below, obtained from 1000 simulations under the null. Using the histogram, approximate (to the best of your abilities) the p-value.



   (e) If instead the consultant actually only had 2 patients who experienced complications, which aspect(s) of your framework above would change and how?

6. In a large university where 60% of the full-time students are employed at least 5 hours per week, the members of the Statistics Department faculty wonder if a higher proportion of their students work at least 5 hours per week. They randomly sample 25 of their majors and find that 20 of the students work 5 or more hours per week.

   Two sampling distributions were created to describe the variability in the proportion of statistics majors who work at least 5 hours per week: a null distribution and a bootstrap distribution. In both cases, $B = 1000$ simulations were generated.



   (a) Describe in words how we could have physically simulated the two different distributions above.

   (b) Using the appropriate histogram, approximate a 90% bootstrap confidence interval for the true proportions of statistics majors who work at least 5 hours per week. Interpret the confidence interval in the context of the problem.

   (c) Using the appropriate histogram, conduct a hypothesis test to answer the faculty's question where the significance level is 0.10. This means you should state the hypotheses, summarise the data, find the p-value, and make a decision in context.

   (d) Briefly comment on how both (c) and (d) might be used to answer the faculty's research question.

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.