

## Homework 4. Due by 11:59pm on Sunday 9/28.

### Data and the reproducibility of research results

Read pages 8–11 of *On Being a Scientist* and [https://en.wikipedia.org/wiki/Data\\_sharing](https://en.wikipedia.org/wiki/Data_sharing). Write brief answers to the following questions, by editing the tex file available at <https://github.com/ionides/810f25>, and submit the resulting pdf file via Canvas.

1. National Institutes of Health (NIH) and National Science Foundation (NSF) require data sharing for research that they fund (excluding infringement of privacy considerations or other individual rights). To what extent do you suspect their rules on data sharing are enforced?

YOUR ANSWER HERE

2. As an academic statistician, you may have the opportunity to work with private data that cannot be shared. What are the advantages and disadvantages of working with unshareable data?

YOUR ANSWER HERE

3. Advanced statistical methods often require sophisticated computational implementations. Should statistical researchers be expected to make their code publicly available (e.g., on GitHub) when they publish results generated using this code?

YOUR ANSWER HERE

The remaining questions consider the following hypothetical case study:

Ben is a Statistics PhD student who has written computer code for a simulation study to test a new statistical theory and methodology which he is developing. He plans to put the results in his thesis and to publish them in a journal paper. The results of the simulations are usually consistent with his theoretical analysis. However, sometimes the code crashes, particularly when investigating more extreme values of the parameter space. Ben has checked and rechecked the code very carefully, and cannot find any error. He asks Gemini to review the code, but that also does not find an error. He decides that there must be some weird numerical effect, perhaps to do with occasional extremely large or small numbers. Ben decides to report the results only in the region of the parameter space where the code never crashed.

4. Is Ben's course of action a reasonable balance between the necessity to make progress on his thesis and his desire to report correct results?

YOUR ANSWER HERE

5. What are the 'data' in this example? What is 'reproducibility' in this context?

YOUR ANSWER HERE

6. Ben asks your opinion on how to proceed. What is your advice?

YOUR ANSWER HERE