**OPTION A: DATA ANALYSIS CASE STUDY**

**Instructions**

1. Submit a compiled PDF, DOCX, or HTML report to Canvas. Excluding code, the report should be at most 900 words. Include all figures in the main text.
2. Submit any code that you used, either as notebooks or a zipped directory.

**Description**

In this assignment, you will have the chance to apply selective inference methods to a real computational genomics dataset. Your report will include three sections: problem formulation, method application, and interpretation/reflection. You can choose a dataset and a method that interests you, but you should do more than just copy an existing data analysis. If you have difficulty identifying a problem, we encourage you to reach out to the instructor. We hope that, through this hands-on case study, you will understand the potential value (but also the limitations) of current selective inference techniques.

In the problem formulation section of your report, describe the scientific question of interest, the data collected, and how your chosen method can bridge the two. Explain why the scientific question is interesting. Summarize the experimental design and the technologies used to generate each dataset. Identify the method you will use to analyze the data. How does it relate to the methods that were discussed during the class sessions? You do not need to explain its algorithmic details, but you do need to justify its applicability to the overall problem.

In the method application section of the report, walk the reader through the code used for data analysis. Describe any preprocessing steps before applying the method, or if the data were already preprocessed when you downloaded it, explain what steps were already taken. Do you have any evidence that the method’s assumptions are met? Did you try running the method with different hyperparameters? Save your results into a separate object/file, so you can return to interpret them without re-running all the methods.

In the interpretation section, generate some tables or visualizations to summarize your results. Describe the main takeaways as if you were writing to a research collaborator. Were the results what you expected, or did you encounter any surprises? Finally, reflect on the overall case study. What aspects of the workflow have the most room for improvement (e.g., computation time, appropriateness of modeling assumptions, the complexity of preparing input data)? Explain your reasoning.

**Rubric**

*Problem Context* [10 points]: Fully researched and developed discussion of the context.

*Implementation* [10 points]: Correct and concise code implementation of the data analysis.

*Interpretation* [10 points]: Accurately and insightfully interprets data analysis results.

**OPTION B: SIMULATION EXPERIMENT**

**Instructions**

1. Submit a compiled PDF, DOCX, or HTML report to Canvas. Excluding code, the report should be at most 900 words. Include all figures in the main text.
2. Submit any code that you used, either as notebooks or a zipped directory.

**Description**

There are bound to be many candidate data analysis approaches for any selective problem. It is often helpful to design benchmarking experiments to help isolate promising directions. By focusing on simple settings where the ground truth is known, these experiments can highlight sensitive and robust methods. This problem gives you a chance to develop such a study. You are free to choose a problem domain of your interest. Your final report should have three main sections: a motivating discussion of the experiment’s context and metrics, details of its design and implementation, and an interpretation of its results.

In the first section, you provide context for your study. What “abstract” selective inference problem that would you like to focus on? What is its scientific motivation, and what types of data are typically available? Describe 2 – 4 techniques that have been proposed for this analysis problem and explain their standard inputs and outputs. Finally, define metrics you could use to compare proposed approaches.

In the second section, you will design and implement a simulation benchmarking experiment comparing the packages you identified above. The requirements are that you (1) compute one of your metrics of interest and (2) vary several parameters that could influence performance. You may either simulate data from scratch or base the experiment on a real dataset – either way, describe your data-generating procedure in detail. Next, apply your selected packages, and run at least one replicate for each setting.

In the final section, you will interpret and reflect on your results. Include at least one figure or table. Was there any noticeable “winner” among the methods you compared? Is the best method context dependent? Did you observe any settings or metrics for which the considered methods are unsatisfactory? What would you recommend in practice, and what are the most interesting opportunities for development?

**Rubric**

*Problem Context* [10 points]: Fully researched and developed discussion of the context.

*Implementation* [10 pointes]: Correct and concise code implementation of the experiment.

*Interpretation* [10 points]: Accurately and insightfully interprets benchmarking results.